UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

Estandarización del protocolo de Captura de Conformación Cromosomal Hi-C e identificación de dominios topológicos en tumores de cáncer gástrico.

TESIS

QUE PARA OPTAR POR EL GRADO DE:
Maestro en Ciencias

PRESENTA:
Biol. Franklin Cruz Villegas

TUTOR PRINCIPAL
Dra. Mayra Furlán Magaril
Instituto de Fisiología Celular, UNAM

MIEMBROS DEL COMITÉ TUTOR

Dra. Erika Patricia Rendón Huerta
Facultad de Medicina, UNAM

Dr. José Carlos Crispín Acuña
Instituto Nacional de Nutrición

Ciudad de México. Agosto, 2023

Agradecimientos

Esta tesis se realizó bajo la dirección y asesoría de la Dra. Mayra Furlan Magaril, en el Departamento de Genética Molecular del Instituto de Fisiología Celular en la Universidad Nacional Autónoma de México.

El Comité Tutoral que siguió y asesoró el desarrollo de este trabajo estuvo formado por:

Dra. Mayra Furlan Magaril	Instituto de Fisiología Celular, UNAM
Dra. Erika Patricia Rendón Huerta	Facultad de Medicina, UNAM
Dr. José Carlos Crispín Acuña	Instituto Nacional de Nutrición

Agradezco el apoyo y asesoría técnica de la Dra. Rosario Pérez Molina, técnica del laboratorio, durante la realización del proyecto y de los experimentos llevados a cabo.

Agradezco a todos mis compañeros del laboratorio por su amistad y su invaluable perspectiva científica. Al “equipo cáncer”, Aura y Eduardo porque con ellos realicé este trabajo y siempre estuvieron dispuestos a enseñarme. Blanca y Andrea que son parte fundamental del proyecto y siempre están atentas.

Al Dr. César Augusto Poot a cargo de la Unidad de Bioinformática y Manejo de la Información IFC, ya que su apoyo fué primordial en este proyecto. Gracias por siempre contestar mis correos y ser muy paciente conmigo.

Agradezco al Programa de Maestría y Doctorado en Ciencias Bioquímicas por apoyar el desarrollo de mi formación científica, al Consejo Nacional de Ciencia y tecnología CONACyT por la beca otorgada con número CVU 1083065 del Programa de Becas Nacionales Para Estudios de Posgrado 2020 proyecto PRONAII 303068.

Agradezco a mi madre, abuela, a mí tía Gloria y a mi familia, esta es la segunda oportunidad que tengo de dedicarles una tesis y nunca estaré suficientemente agradecido. Saben que son parte de mi, más allá de la Biología, su educación, valores, sacrificio y amor son los factores que me han llevado a lograr todo esto, ustedes me hacen fuerte.

A mis amigos, Alejandro, Alberto, Brandon, Anahí y Priscila, por su apoyo y consejo, los admiro profundamente.

A mi tutora la Dra. Mayra Furlan Magaril, gracias por su guía, paciencia y por compartirme sus conocimientos, por haberme dado la oportunidad de realizar mi primer trabajo científico en Epigenética. Es un ejemplo a seguir.

Abreviaturas

- ARG- *Asian Cancer Research Group*
- CG- Cáncer Gástrico
- CIN- Inestabilidad Cromosómica
- CNV- Copy Number Variation
- D- Digerido
- EBV- Epstein-Barr Virus
- EMT- Transición Epitelio Mesénquima
- FISH- Hibridación Fluorescente *in situ*
- FT- Factores de Transcripción
- GDE- Genes Diferencialmente Expresados
- GS- Genómicamente Estable
- Hi-C- Captura Conformacional Cromosomal de alto rendimiento
- Kb- Kilobase
- Mbp- Megabase
- MSI- Microsatélites Inestables
- MSS- Estabilidad de Microsatélites
- ND- No Digerido
- PARP-Inhibidor de la PolyADP-Ribosa
- PCA- Análisis de Componentes Principales
- PMBCs- Peripheral Mononuclear Blood Cells
- ScRNA-seq- Single Cell RNA- seq
- SNPs – Polimorfismos de nucleótido único
- TADs- Dominios Topológicamente Asociados
- TCGA- *The Cancer Genome Atlas*
- TFBS- Sitios de Unión de Factores de Transcripción
- t-SNE- T- Distributed Stochastic Neighbor Embedding
- UMAP- Uniform manifold aproximation and Project
- VE- Variaciones Estructurales

“No hacemos ciencia para el pueblo, somos el pueblo haciendo ciencia.”

Pinta en la Facultad de Ciencias.

“ Valga la comparación: Colón tuvo un viaje peor que Lindberg, después de todo; y los buenos científicos hacen cosas cuya primera medida de dificultad es que nadie las ha hecho antes. ”

El Octavo día de la Creación por H.F. Hudson.

“Computing should be no more mystical a technique than any other laboratory method”

Computer Applications in the Biosciences, 1985.

Contenido

Resumen

En 2020 hubo más de 1, 000, 000 de nuevos casos de cáncer gástrico (CG) a nivel mundial y se estima que provocó la muerte de 769, 000 personas ese mismo año. Representa el quinto tipo de cáncer más diagnosticado, la tercera causa principal de muerte por cáncer en México y la cuarta a nivel mundial[1,2].

El cáncer gástrico del subtipo difuso afecta a personas más jóvenes, es más agresivo y la sobrevida esperada es menor[3–5]. Además de que la incidencia de esta enfermedad se ha relacionado con poblaciones hispanas[6]. No existen tratamientos específicos para el cáncer gástrico del subtipo difuso. Por lo tanto, es indispensable conocer a nivel molecular esta patología en pacientes mexicanos ya que esto puede ayudar al diseño de nuevos tratamientos dirigidos.

En este trabajo se estandarizó la técnica de captura conformacional cromosomal de alto rendimiento “Hi-C” por primera vez, en células mononucleares de sangre periférica (PMBCs), tejido adyacente sano y tumor de cáncer gástrico del subtipo difuso de un paciente mexicano. Además, se estandarizaron flujos de trabajo bioinformáticos “pipelines” que permiten la detección de variaciones estructurales (VE) en matrices de Hi-C y se demostró su eficiencia al comparar los datos obtenidos con una biblioteca de Hi-C de CG del subtipo intestinal y otra derivada de una línea celular de CG, previamente publicadas[7]. Los resultados apoyan la hipótesis de que el cáncer gástrico difuso es en general genómicamente estable no así, el subtipo intestinal[4,8,9]. El análisis y comparación de estos datos, demuestra que el Hi-C es muy eficiente para la detección de VE incluso cuando se dispone de muestras de bajo número celular, como es el caso común de las biopsias de CG.

Además se encontró que a nivel de dominios topológicamente asociados (TADs), que el tejido adyacente sano y el tumor de CG del subtipo difuso comparten solamente cerca de la mitad de TADs, mostrando un cambio topológico global en las células del tumor, esta observación es apoyada por la integración de datos transcriptómicos de scRNA-seq donde se detectaron 1811 genes diferencialmente expresados (GDE) de los cuales 717GDE se ubican en TADs exclusivos del tumor.

Al realizar el análisis de enriquecimiento se mostró que el proceso biológico más relevante con el que están relacionados los GDE es el cargado de la cohesina a la cromatina, estando sobreexpresados en el tumor todos los genes que forman el núcleo del complejo de la cohesina (*SMC3, SMC1A, RAD21 y STAG2*), los genes que regulan su cargado (*WAPL y PDS5A*) y descargado (*NIPBL*) de la cromatina. Debido a la importancia de este complejo en la topología del genoma se indagó en los TADs que contienen a estos genes, observando que 6 de los 7 genes se encuentran en TADs exclusivos del tejido tumoral, siendo *NIPBL* el que comparte TAD con el tejido adyacente sano. Por lo que se sugiere a estos *loci* como candidatos para estudiar la regulación genética del CG del subtipo intestinal.

Además, esta característica podría estar relacionada con la estabilidad genómica de la malignidad, ya que el complejo de la cohesina está implicado en la reparación de la ruptura de doble cadena del ADN aunado a que se ha reportado anteriormente la sobrexpresión de las subunidades *SMC1A* y *RAD21* en tumores de CG[10], particularidad que corresponde con la resistencia a fármacos de daño al ADN como el Cisplatino e inhibidores de PolyADP-Ribosa (PARP), ambos utilizados en el tratamiento general del CG. Por lo tanto existe la posibilidad de que los mismos no sean los más adecuados para tratar el CG del subtipo difuso, acercándonos a la búsqueda de nuevas propuestas terapéuticas.

Finalmente, estas observaciones representan un primer paso en México para la integración de datos topológicos y transcripcionales en el cáncer gástrico, es importante recalcar que son necesarios más datos para robustecer o rechazar estas propuestas, parte de este trabajo también implicó el inicio de una colección de muestras de esta patología con las cuales se seguirá trabajando con experimentos de Hi-C y scRNA-seq.

Abstract

In 2020, there were over 1,000,000 new cases of gastric cancer (GC) worldwide, resulting in an estimated 769,000 deaths. It represents the fifth most diagnosed cancer, the third leading cause of cancer-related deaths in Mexico, and the fourth worldwide.

The diffuse subtype of gastric cancer affects younger individuals, is more aggressive, and has a lower expected survival rate. Additionally, its incidence has been associated with Hispanic populations. Currently, there are no specific treatments for the diffuse subtype of gastric cancer. Therefore, it is essential to understand this pathology at a molecular level in Mexican patients to aid in the development of targeted therapies.

This work successfully standardized the high-throughput chromosomal conformation capture technique, known as "Hi-C," for the first time in peripheral blood mononuclear cells (PMBCs), adjacent healthy tissue, and diffuse subtype gastric cancer tumor from a Mexican patient. Bioinformatic pipelines were also established to detect structural variations (SV) in Hi-C matrices, and their efficiency was demonstrated by comparing the data with a library of Hi-C from intestinal subtype GC and a CG cell line, previously published. The results support the hypothesis that diffuse gastric cancer is generally genomically stable, unlike the intestinal subtype [4,8,9]. The analysis and comparison of these data demonstrate that Hi-C is highly efficient in detecting SV, even with samples of low cell numbers, as is often the case in CG biopsies.

Furthermore, at the level of topologically associated domains (TADs), it was found that adjacent healthy tissue and diffuse CG tumor share only about half of the TADs, indicating a global topological change in tumor cells. This observation is supported by the integration of transcriptomic data from scRNA-seq, where 1811 differentially expressed genes (DEGs) were identified, with 717 DEGs located in TADs exclusive to the tumor.

Enrichment analysis revealed that the most relevant biological process associated with DEGs is the loading of cohesin onto chromatin, with all core cohesin complex genes *(SMC3, SMC1A, RAD21*, and *STAG2*), as well as genes regulating cohesin loading (*WAPL* and *PDS5A*), and cohesin unloading (*NIPBL*) being overexpressed in the tumor. Due to the importance of this complex in genome topology, further investigation was conducted into TADs containing these genes, revealing that 6 out of 7 genes are found exclusively in tumor TADs, while *NIPBL* shares a TAD with adjacent healthy tissue. This suggests these loci as candidates for studying the genetic regulation of diffuse CG.

Additionally, this characteristic may be related to genomic stability in malignancy since the cohesin complex is involved in double-strand DNA break repair. Prior reports have shown overexpression of SMC1A and RAD21 subunits in GC tumors, which correlates with resistance to DNA-damaging drugs such as Cisplatin and PolyADP-Ribose (PARP) inhibitors, both used in the general treatment of GC. Hence, there is a possibility that these drugs may not be the most suitable for treating the diffuse subtype of CG, leading to the search for new therapeutic approaches.

In conclusion, these observations represent a crucial step in Mexico towards integrating topological and transcriptional data in gastric cancer. It is important to emphasize that more data is needed to further validate or reject these proposals. Furthermore, this work has initiated the collection of samples from this pathology, with ongoing experiments of Hi-C and scRNA-seq.

Introducción

En 2020 hubo más de 1, 000, 000 de nuevos casos de cáncer gástrico a nivel mundial y se estima que provocó la muerte de 769, 000 personas ese mismo año. El cáncer gástrico representa el quinto tipo de cáncer más diagnosticado, la tercera causa principal de muerte por cáncer en México y la cuarta a nivel mundial[1,2].

Algunas áreas geográficas de la República Mexicana tienen mayores tasas de mortalidad por cáncer gástrico en contraste con otras como por ejemplo Chiapas, cuya tasa es de 6.4 por cada 100,000 habitantes en comparación con la Ciudad de México con 4.5 y el Estado de México con 2.5[11]. Esto sugiere que, además de la influencia de variables ambientales como la alimentación, el consumo de alcohol, tabaco e infección por *H. pylori,* entre otras, podría existir un componente genético subyacente de predisposición a desarrollar cáncer gástrico en la población mexicana [5].

Clasificación histológica del cáncer gástrico

Existen distintas clasificaciones del cáncer gástrico. La clasificación histológica de Lauren, es la más utilizada en clínica y distingue dos tipos, el intestinal y el difuso [12] (Figura 1). El tipo intestinal está asociado con gastritis atrófica crónica y metaplasia intestinal y se presenta más a menudo en el estómago distal. El tipo difuso se origina normalmente en la mucosa gástrica, presenta células características denominadas “de anillo de sello” las cuales contienen una gran vacuola llena de mucosa. Este subtipo es común en personas menores de 45 años y es el tipo de cáncer gástrico con peor pronóstico. Es decir, la sobrevida a 5 años después del diagnóstico de la enfermedad no supera el 10%[3]. Cerca del 3% del cáncer gástrico difuso es hereditario debido a mutaciones en el gen *CDHI* que codifica para la E-cadherina, una proteína importante en la adherencia celular[5,13,14]. Sin embargo, la heterogeneidad genética en el cáncer gástrico del tipo difuso está poco estudiada.

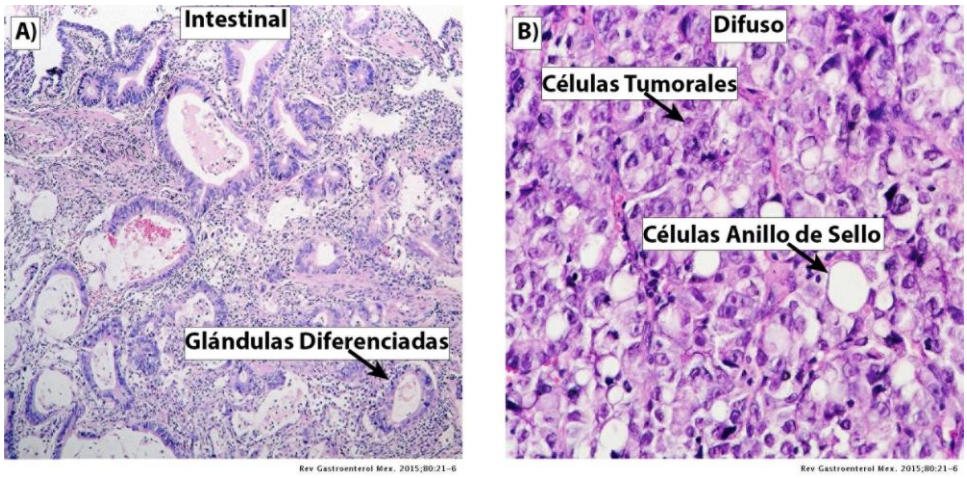


Figura 1. Tipos de cáncer gástrico según la clasificación de Lauren. A) Glándulas diferenciadas características en la histología del tipo intestinal. **B)** Células tumorales con grandes núcleos y las células de anillo de sello presentando una gran vacuola llena de mucosa típicas del tipo difuso (Modificada de Martínez-Galindo *et al.* 2015[12]).

Clasificación molecular

Existen otros esfuerzos de clasificación a nivel molecular de esta patología derivados de las tecnologías de secuenciación masiva paralela. Dos de los más recientes incluyen al estudio del *The Cancer Genome Atlas* (TCGA por sus siglas en inglés) publicado en 2014 por Bass *et al*[9]. donde realizaron una caracterización molecular integral de 295 adenocarcinomas gástricos primarios. A partir de esta caracterización propusieron un nuevo sistema de clasificación que comprende cuatro subtipos: 1. Tumores positivos para Virus de Epstein-Barr (EBV). Estos presentan amplificaciones de los genes *PD-L1/2* (relacionados con la evasión del sistema inmune) y *JAK2* que promueve la reproducción celular y son negativos para las mutaciones en *TP53*. 2. Tumores con microsatélites inestables (MSI) que presentan hiper mutación en *TP53*. Estos son más comunes en personas de mayor edad. 3. Tumores genómicamente estables (GS) que regularmente tienen una histología difusa, son más comunes en personas jóvenes y contienen mutaciones en genes relevantes en la adhesión celular y 4. Tumores con inestabilidad cromosómica (CIN) relacionados con hipermutación de *TP53* e histología intestinal (Figura 2) [8,9].

Por otro lado, en 2015 el *Asian Cancer Research Group* (ACRG), Analizó datos de expresión génica de 300 tumores gástricos primarios. Sus hallazgos han llevado a una propuesta de clasificación molecular del cáncer gástrico que incluye cuatro subtipos de tumores: 1. Tumores con estabilidad de microsatélites y transición epitelio mesénquima (MSS/EMT), presentan pérdida de *CDH1* e histología difusa. 2. Tumores con microsatélites inestables (MSI) con sobreexpresión de *PD-L1*. 3.(MSS/P53+) con microsatélites estables y *TP53*-activo, presentan histología intestinal y 4. Tumores con *TP53* inactivo presentando microsatélites estables (MSS/TP53$^-$) con histología intestinal y mutaciones en *TP53* (Figura 2)[4,8].

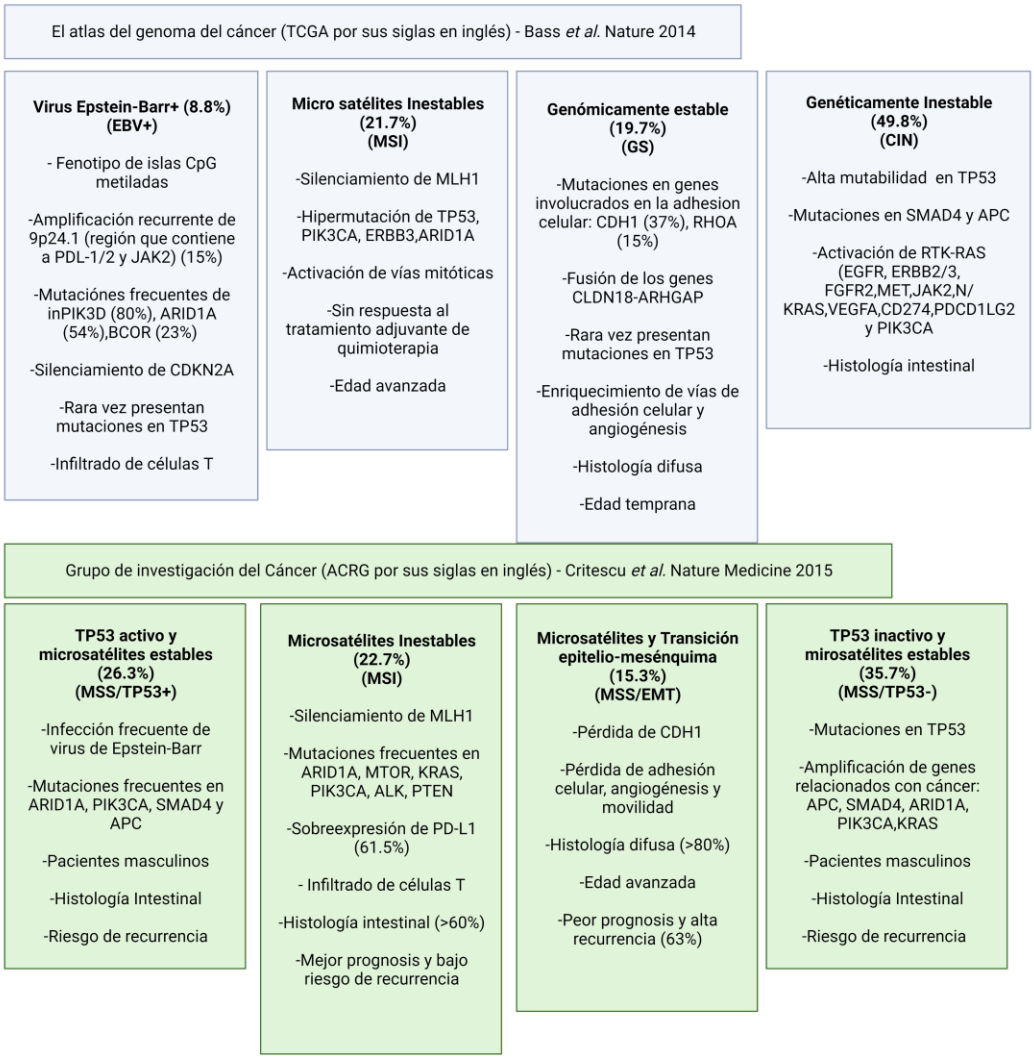


Figura 2. Características relevantes en la clasificación molecular del cáncer gástrico. Distinción de parámetros utilizados para clasificar molecularmente a los distintos tumores de cáncer gástrico por los grupos de TCGA y ACRG. Cabe resaltar que se muestra el porcentaje de cada subtipo de tumor gástrico en la muestra de tejidos analizados por cada grupo (Modificada de Chivu-Economescu et al. 2018[8]).

Las clasificaciónes existentes del cáncer gástrico nos muestran la enorme heterogeneidad molecular que subyace a esta enfermedad y de como la pobre prognosis de este cáncer se debe en alguna medida a la falta de tratamientos dirigidos. En ambos tipos histológicos encontramos tratamientos generales, siendo la cirugía la única terapia curativa, mientras que la quimioterapia perioperatoria y adyuvante, así como la quimiorradiación, pueden mejorar el resultado de la resección de este cáncer, además de la disección de los ganglios linfáticos adyacentes[3,5].

Sin embargo, más de la mitad de los pacientes con cáncer gástrico resecado recaen con tumores locales o con metástasis, o reciben el diagnóstico de cáncer gástrico cuando el tumor se disemina. Esto hace que la media de supervivencia global a esta enfermedad rara vez supere los 12 meses y la supervivencia a los 5 años sea inferior al 10%[3]. Estos hechos hacen inminente el estudio a profundidad de estos tipos tumorales en nuestra población para comprender mejor no solo la genética del cáncer gástrico, si no también los mecanismos de regulación transcripcional que se encuentren alterados, contribuyendo así al descubrimiento de nuevos marcadores de diagnóstico y sobre todo al diseño de tratamientos más eficientes y dirigidos.

Topología del Genoma

El ADN en interfase se empaqueta dentro del núcleo celular formando estructuras a distintas escalas. El material genético se pliega en nucleosomas que a su vez forman asas de cromatina. A escalas de centenas de kilobases se forman estructuras denominadas dominios topológicamente asociados (TADs por sus siglas en inglés), a escalas de megabases se forman compartimentos y finalmente los cromosomas completos se ubican en territorios cromosómicos[15].

El desarrollo de tecnologías de secuenciación masiva paralela ha propulsado nuestra capacidad de caracterizar funcionalmente diferentes genomas. En particular, las estrategias de Captura Conformacional de Cromosomas como el Hi-

C posibilitan la reconstrucción de mapas de contactos genómicos a nivel global permitiendo la detección de asas de cromatina, TADs y compartimentos presentes en el genoma de la población celular en alta resolución[16,17]. Así mismo, el Hi-C permite detectar todas las variaciones estructurales (VE) mayores o iguales a 1Mb presentes en el genoma sin conocimiento previo de las mismas[18].

La técnica de Hi-C se basa en varios principios clave. En primer lugar, se realiza el fijado de las células con formaldehído para estabilizar las interacciones espaciales entre regiones genómicas. Luego, se lleva a cabo la digestión de la cromatina, que implica la fragmentación del ADN mediante enzimas de restricción. Posteriormente, las moléculas fragmentadas se ligan para formar un complejo circular, que permite la detección de las interacciones espaciales. A continuación, se realiza la biotinilación de las moléculas de ADN, lo que facilita su posterior enriquecimiento y secuenciación. Este proceso permite analizar la estructura tridimensional del genoma y revelar la organización espacial de los cromosomas[15,19,20]. El protocolo para realizar esta técnica se describirá con detalle en la sección de materiales y métodos.

Las matrices de Hi-C o mapas de contactos genómicos se interpretan teniendo en cuenta que la frecuencia de interacción entre *loci* disminuye a medida que aumenta la distancia genómica entre ellos, se pueden visualizar las matrices como mapas de calor, donde los píxeles de la matriz representan la frecuencia de interacción entre dos regiones genómicas. Las regiones con una alta frecuencia de interacción se muestran en tonos más intensos y las de menor frecuencia en tonos más claros[19,20]. A continuación se describirán las estructuras biológicas detectadas por Hi-C y sus implicaciones en la regulación genética.

Asas de Cromatina

Las asas de cromatina son relevantes en el genoma de los vertebrados ya que a pesar de la lejanía que puede existir entre *loci* en el genoma lineal, las asas ponen en proximidad espacial a elementos de regulación genómicos con los promotores de sus genes diana. Para que dos *loci* genómicos distantes interactúen dentro de una escala de tiempo funcionalmente relevante, la cromatina necesita ser compactada en una escala promedio de longitud genómica de 10kb-1Mb y explorar su entorno dinámicamente[17,21,22].

Por ejemplo, en el genoma de las células troncales embrionarias humanas hay al menos 13.000 asas de cromatina, que varían de 25 kb a 940 kb de tamaño y contienen de uno a diez genes[23]. La longitud media de una asa de cromatina es ~190 kb y contienen en promedio tres genes[23–25].

Las asas de cromatina están formadas por la interacción de homodímeros del factor de unión CCCTC (CTCF) estabilizados por cohesina. Se ha demostrado que entre el 70-85% de las asas de cromatina son delimitadas por CTCF, dependiendo del tipo celular[20,26] y ~80% de estos sitios frontera fueron también ocupados por el complejo de cohesina[24,27] (Figura 3).

Estas observaciones tienen una estrecha relación con el funcionamiento de la formación de las asas de cromatina y la regulación de la expresión genética. El modelo de extrusión de las asas de cromatina propone que estas se forman por la actividad de factores de extrusión como las cohesinas, que progresivamente extruyen grandes regiones de cromatina de manera ATP dependiente hasta que se disocian de la cromatina o encuentran una frontera. Por ejemplo, los dímeros de CTCF[24].

El núcleo del complejo de la cohesina está compuesto por cuatro subunidades SMC1A, SMC3, RAD21 y STAG2 (antígeno estromal) la asociación de las mismas genera una estructura en forma de anillo[28] (Figura 3).

SMC1A y SMC3 están configuradas cada una por una hélice antiparalela superenrollada que en un extremo presenta un dominio globular flexible que al heterodimerizarse una con la otra forman un dominio bisagra en forma de V que se mantiene por interacciones hidrófobas. En el lado contrario están el extremo N-terminal que contiene un motivo Walker A que se une al ATP durante la extrusión de las asas de cromatina y el extremo C-terminal con el motivo Walker B que se asocia con el ADN[28]. RAD21 interacciona con SMC1A en su externo carboxilo y con SMC3 en su extremo amino en estas regiones y cierra el anillo, mientras STAG2 es esencial para la asociación del complejo con el ADN[28–30] (Figura 3).

La regulación del cargado del complejo de cohesina a la cromatina está dado por la proteína NIPBL (Nipped-B like) y está relacionado con la concentración de la misma en el núcleo afectando la expresión genética[31]. Además, PDS5A (proteína de disociación precoz de cromátidas hermanas 5) participa en el proceso manteniendo a la cohesina asociada a la cromatina. Por otro lado, la disociación del anillo de cohesina asociada a la cromatina requiere la participación de WAPL (proteína Wings apart-like) (Figura 3), modelos de silenciamiento de WAPL han generado una organización del genoma en grandes loops y se han asociado con desregulación de la expresión genética[28–31].

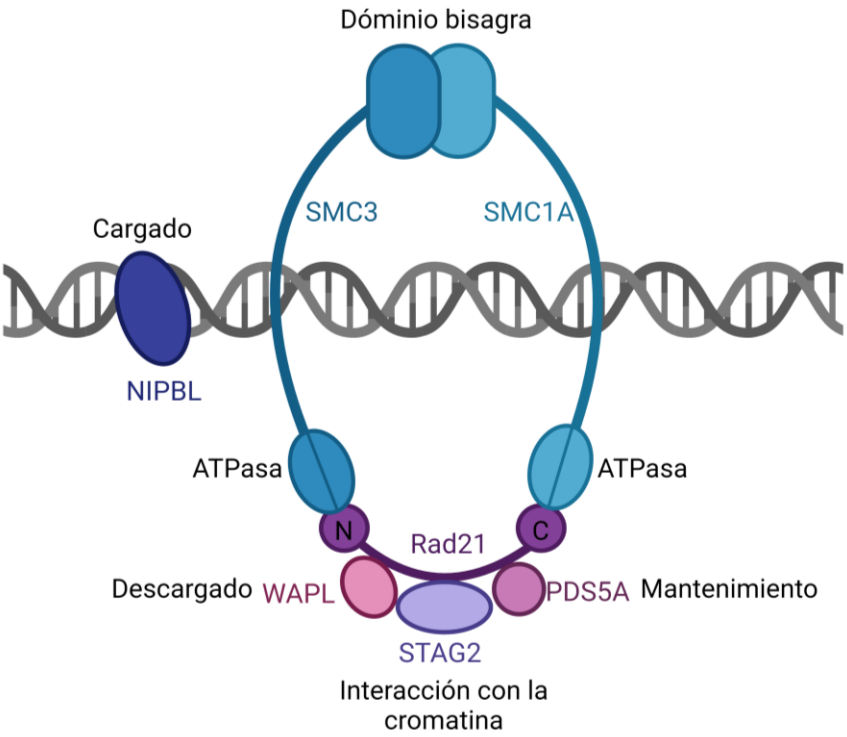


Figura 3. Modelo del complejo de la Cohesina. Se muestran las subunidades del nucleo de la cohesina y proteínas reguladoras con sus funciones asociadas (Modificada de Losada, 2014[29]).

Se sabe que la formación de los dímeros de CTCF solo sucede cuando los motivos de unión de esta proteína en el genoma se encuentran de forma convergente, lo cual sugiere una disposición de las asas de cromatina codificada en el genoma[20,27].

Por lo tanto, las variaciones específicas de la topología del genoma de las células de cada tejido y su expresión, se podrían explicar a través de la expresión diferencial de los reguladores que establecen la velocidad de la extrusión de las asas por el complejo de la cohesina, la eficiencia de carga del extrusor, concentración del extrusor, la fuerza de unión y polaridad de las proteínas de frontera como CTCF. Es decir, proteínas que impidan la extrusión de la cromatina por el complejo de la cohesina[24].

Un ejemplo de elementos de regulación distal en el genoma son los amplificadores de expresión o “enhancers” que son secuencias que activan la expresión de genes diana transcritos por la ARN polimerasa II (ARNPII) al contactar con su promotor y reclutar maquinaria transcripcional, todo esto en un contexto de asa de cromatina (Figura 4 A).

Los potenciadores pueden actuar independientemente de su orientación, distancia y ubicación con respecto al gen objetivo[32] y se pueden ubicar a más de un millón de pares de bases de distancia, como se observa en el caso del gen SHH[33] en células humanas T CD4+ el cual interacciona con un potenciador llamado ZRS que se encuentra aproximadamente a 1 Mb del gen[34,35].

Los potenciadores de vertebrados tienen una longitud de entre 100 a 1000 pb y pueden existir múltiples potenciadores en un grupo para formar un súper potenciador[36,37]. Los potenciadores se encuentran principalmente en regiones intergénicas y regiones intrónicas, Sin embargo, se han encontrado algunos dentro de los exones[32].

Los potenciadores consisten en grupos densos de sitios de unión de factores de transcripción (TFBS por sus siglas en inglés) y están unidos a factores de transcripción (FT) específicos del tipo celular donde se encuentran, correguladores, modificadores de cromatina, proteínas arquitecturales como cohesina, condensina, CTCF y ARNPII. Cuando un enhancer está activo participan más de 45 proteínas teniendo una masa combinada aproximada de 2500 kDa[32].

Debido al ensamblaje de tantas proteínas, a menudo cuando están activos son deficientes en nucleosomas y, por lo tanto, son hipersensibles a las nucleasas, característica relacionada con la accesibilidad del ADN y ampliamente explotada para identificar potenciadores[38]. Además, se han relacionado las marcas de histonas como la monometilación de la lisina 4 de la histona 3 (H3K4me1) común a todos los potenciadores y la acetilación de la lisina 27 de la histona 3 (H3K27ac)

para identificar potenciadores activos[39]. Una vez ensamblado, el complejo del potenciador contacta el promotor objetivo y activa la transcripción.

Un caso de la acción de los potenciadores es el *locus* de la región de control (LCR) del grupo de β globina, que interactúa a través de contactos de cromatina de largo alcance, con sus genes diana en las células eritroides (donde el gen de la ***β-globina*** está activo) pero muestra poca o ninguna interacción en células de otros linajes como células troncales o neuronales[15,40]. Mostrando que las asas de cromatina a menudo son tipo celular dependientes. En cáncer gástrico, el secuestro de un potenciador provocado por la duplicación en tándem de un *locus* del cromosoma 19 genera la sobreexpresión del gen *CCNE1*, una ciclina relacionada con proliferación y menor sobrevida de los pacientes con esta malignidad[7].

Por último, es importante señalar que las asas de cromatina no se limitan solo a interacciones “potenciador-promotor” también existen asociaciones espaciales entre genes co-regulados que se transcriben activamente compartiendo maquinaria y entre genes reprimidos por Polycomb tanto en *Drosophila melanogaster* como en mamíferos[15,41–43] (Figura 4 A).

Dominios topológicamente asociados, TADs

Los conjuntos de asas de cromatina se organizan en TADs cuya longitud media es de 880 Kb[44] (Figura 4 B y C). La mayoría de los TADs son estables entre tipos celulares y están aislados entre sí por fronteras delimitadas por dímeros de CTCF y cohesina[45] (Figura 4 B, Ver Figura 3). Existen alrededor de 2000 TADs en el genoma del ratón con un promedio de 8 genes en su interior[25,44]. En tanto al genoma humano se han detectado alrededor de 3741 TADs en células de sangre periférica[46].

Además de CTCF y cohesina, las fronteras de los TADs se caracterizan por marcas de histonas asociadas a transcripción activa como la trimetilación de la lisina 4 de la histona 3 (H3K4me3) y la trimetilación de la lisina 36 de la histona 3 (H3K36me3)

y a elementos repetidos[15]. Al interior de los TADs se presenta generalmente un estado transcripcional uniforme[15,24].

A menudo en las matrices de Hi-C se presentan TADs anidados o Sub-TADs que corresponden a asas de cromatina más pequeñas dentro de los TADs que generalmente tienen puntajes de aislamiento más débiles, marcas de cromatina más homogéneas y exhiben mayor heterogeneidad entre las células[15,24].

Dentro de los TADs las interacciones cromosómicas se reorganizan dinámicamente respecto a distintos procesos celulares como la diferenciación o la carcinogénesis. En experimentos donde se depletan las fronteras de los TADs se ha encontrado una pobre disminución en la expresión genética mostrando evidencia de que estas estructuras tienen funciones de aislamiento para las secuencias dentro de ellas evitando interacciones ectópicas, pero que no regulan la expresión génica por sí mismos[44,47].

En resumen, los TADs son importantes estructuras tridimensionales en el genoma, que organizan la cromatina en regiones funcionales. Su delimitación por proteínas como CTCF y cohesina, junto con las marcas de histonas asociadas, contribuye a mantener la estabilidad y la regulación adecuada de la expresión génica. La presencia de Sub-TADs agrega un nivel adicional de complejidad en la arquitectura cromosómica, permitiendo la adaptación a distintos estados celulares. Aunque los TADs están involucrados en el mantenimiento de la integridad estructural del genoma, su papel exacto en la regulación génica sigue siendo objeto de investigación para comprender plenamente su contribución a los procesos biológicos y su relevancia en diversas enfermedades.

Compartimentos de cromatina

Los compartimentos de cromatina son estructuras observadas en las matrices obtenidas por Hi-C, estos se caracterizan en la matriz por tener un patrón de "tablero de ajedrez" donde cada cuadrante es de varias Mb de longitud[18,19] y están divididos en compartimentos A y B que corresponden a eucromatina y heterocromatina respectivamente[48] (Figura 4 C). Las interacciones entre regiones del genoma del mismo compartimento son más frecuentes y existen pocos contactos entre regiones genómicas de compartimentos diferentes. Además, los compartimentos se localizan en diferentes zonas del núcleo. El compartimento A se encuentra preferentemente al interior del núcleo mientras que el B está ubicado cerca de la periferia nuclear y se asocia con las laminas nucleares, formando amplias regiones de heterocromatina[49,50]. A diferencia de los TADs que están muy conservados a lo largo de los tipos celulares los compartimentos son específicos del tipo celular[15].

Territorios cromosómicos

Finalmente, los cromosomas ocupan espacios discretos en el núcleo llamados territorios cromosómicos. Identificados originalmente por técnicas como la hibridación fluorescente *in situ* (FISH por sus siglas en inglés)[51] y actualmente estudiados por Hi-C, muestran que las interacciones dentro de cada cromosoma (*cis*) son más frecuentes que entre cromosomas (*trans*)[15,19] (Figura 4 D).

En conjunto, la topología del genoma está estrechamente ligada con la regulación genética y su estudio muestra con detalle los mecanismos por los cuales los genes se expresan, debido a la importancia de la localización en el espacio de las regiones genómicas. Si bien el perfil de expresión genética de muchos tipos de cáncer ha sido ampliamente estudiado, el componente topológico que explica la expresión no ha sido explorado con detalle. Particularmente en cáncer gástrico solo existe un artículo que reporta experimentos de Hi-C en tumores del subtipo intestinal[7]. Por lo tanto es importante seguir indagando en la topología genómica de esta malignidad.

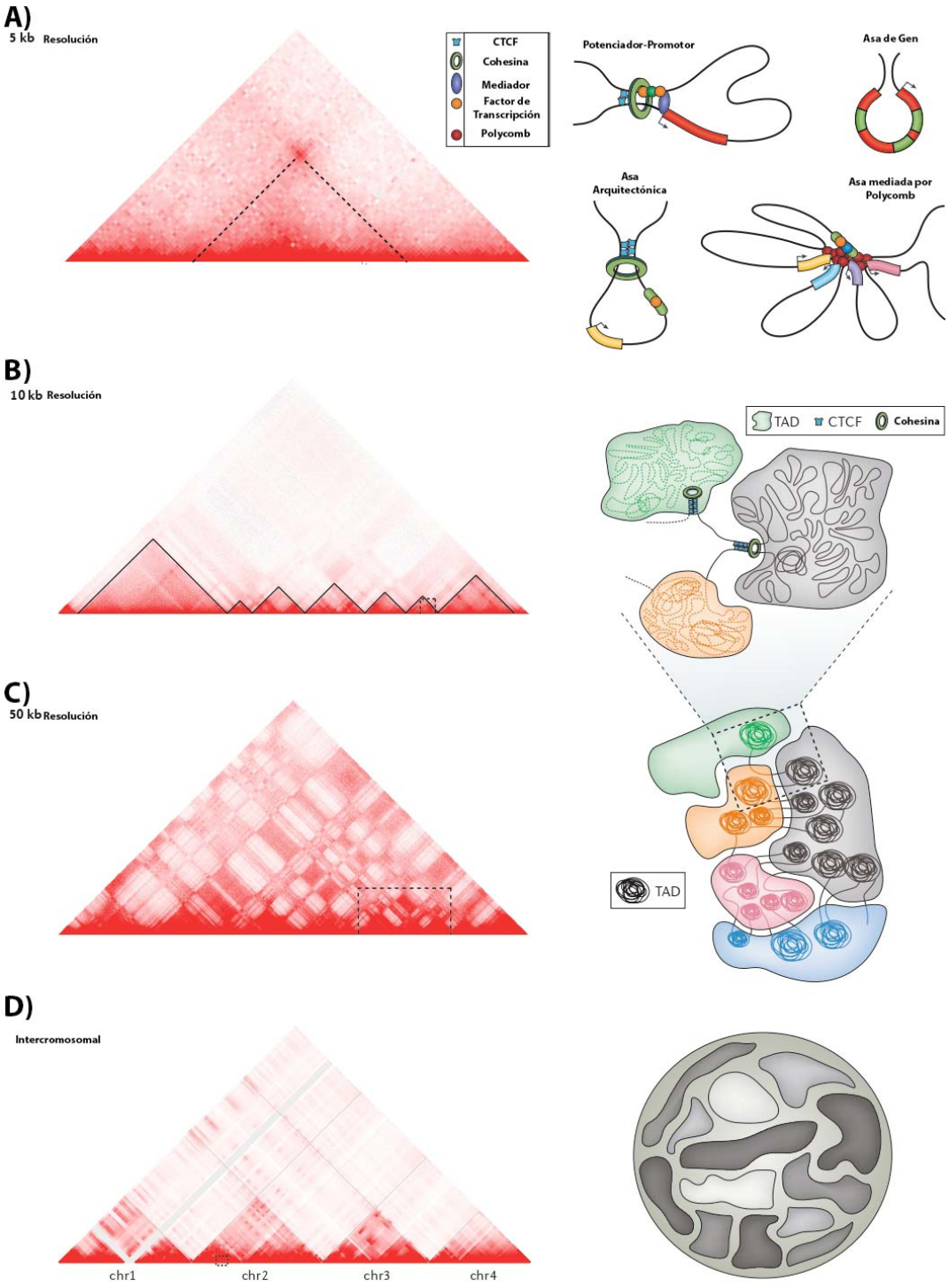


Figura 4. Los distintos niveles de organización del genoma. A) Modelos de la interacción entre el potenciador y el promotor de un gen en un asa de cromatina, asa formada por un solo gen y asas silenciadas por polycomb. **B)** Formación de TADs delimitados por dímeros de CTCF y Cohesina, **C)** Formación de compartimentos y **D**) Territorios cromosómicos. En todos los casos a la izquierda se muestran las matrices de Hi-C correspondientes a cada estructura con la resolución a la que se pueden observar (Modificada de Bonev & Cavalli, 2016[15]).

Antecedentes

Arquitectura del genoma en el cáncer

La inestabilidad genómica y en consecuencia los rearreglos cromosómicos, son características distintivas del cáncer. Fallas en la maquinaria molecular de reparación y rompimientos de la doble cadena de ADN conllevan a la formación de aberraciones cromosómicas o VE[52]. La desregulación de los niveles transcripcionales de genes diversos es uno de los mecanismos por el cual las células tumorales obtienen una alta capacidad proliferativa y de evasión del sistema inmune produciendo metástasis.

Se ha reportado que cerca del 40% de todos los tumores malignos presentan mutaciones en p53 que afectan la capacidad de reparación del material genético en las células afectadas[53]. También se ha demostrado que muchos tipos de cáncer hereditario contienen defectos en la maquinaria de reparación aumentando la posibilidad de errores mutagénicos durante la replicación del DNA[53].

El desarrollo de técnicas de microscopía como la hibridación fluorescente *in situ* (FISH por sus siglas en inglés) y la secuenciación masiva ha confirmado que casi todas las formas de cáncer involucran células cuyos genomas han sido alterados ya sea a través de rearreglos cromosómicos complejos, mutaciones puntuales o una combinación de ambas[53–56].

Las VE existen en dos formas generales: Balanceadas y No-Balanceadas. Las balanceadas o neutrales en el número de copias, se refieren a aquellas en las que no hay ganancia o pérdida neta de material genético, aunque se ha producido algún reordenamiento. Estas incluyen a las translocaciones e inversiones recíprocas (Figura 5 A,C). En las VE desequilibradas existe pérdida o ganancia de material genético, por ejemplo las deleciones, duplicaciones y la herencia de productos de translocación desequilibrada (Figura 5 B,D,E).

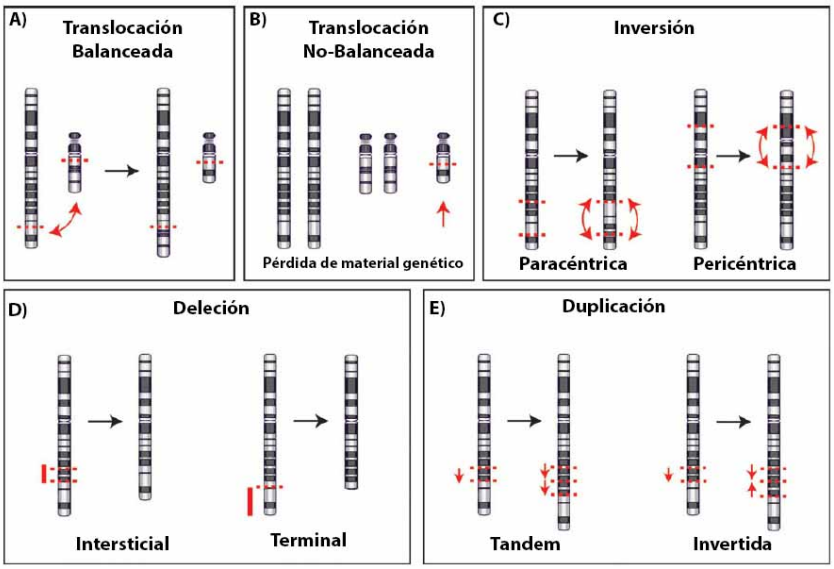


Figura 5. Ejemplos de Variaciones Estructurales (VE). A) Translocaciones equilibradas sin ganancia o pérdida neta de material genético. **B)** Un cromosoma derivado adicional (flecha roja) que se ha heredado con una translocación desequilibrada resultando en trisomía de regiones del cromosoma dañado. **C)** Inversiones, en las que se invierte un segmento de un cromosoma, pueden resultar de dos puntos de ruptura (líneas rojas punteadas) situadas en un brazo cromosómico (inversión paracéntrica) o en ambos lados del centrómero (inversión pericéntrica). **D)** Deleciones entre puntos de ruptura en los brazos del cromosoma (deleción intersticial) o en los extremos del cromosoma (deleción terminal). **E)** Duplicaciones. Estas forman dos categorías, dependiendo de si los segmentos duplicados se encuentran en la misma orientación (duplicación en tándem) o se han volteado (duplicación invertida), (modificada de Harewood & Fraser, 2014[57]).

Las VE pueden afectar zonas codificantes del genoma de modo que alteran directamente la estructura de un gen y la proteína a la que codifica. Por ejemplo, la translocación t(9;22) (q34;q11) en leucemia mieloide crónica que causa la fusión de los genes *BCR-ABL* dando un cromosoma 22 alterado denominado cromosoma filadelfia y cuya expresión inhibe la reparación del ADN[57,58]. En CG difuso una translocación balanceada entre los cromosomas 3 y el 5 provoca la fusión del gen de *Cldn18* un proteína de uniones estrechas que mantiene la integridad epitelial y el inhibidor de RHOA, *ARHGAP26* generando transcritos aberrantes o fusionados, la expresión de estos transcritos se ha relacionado con aumento en la proliferación celular y con el proceso de transición epitelio mesénquima (TEM por sus siglas en inglés) debido a la pérdida de uniones estrechas[59,60].

En muchas ocasiones, las VE también pueden encontrarse en regiones no codificantes del ADN, afectando elementos de regulación distales o la organización tridimensional del genoma. Esto puede llevar a la desregulación de la expresión de genes. Un ejemplo de esto es observado a nivel epigenético en el linfoma de células B, donde la translocación desequilibrada (+der(2)t(1;2)(q12;p13)) muestra cambios en la cromatina de las secuencias cercanas al punto donde se integró en el cromosoma 2p. Como resultado de esta translocación, estas regiones adquieren marcas epigenéticas represivas, como H4K20me3 y H3K9me3 [57,61], que afectan la expresión genética del *locus*.

Las VE pueden tener un impacto significativo en la regulación génica y la expresión de genes al alterar la organización tridimensional del genoma o afectar elementos de regulación distales. En la actualidad se cuenta con muy poca información con respecto a las VE existentes en genomas de células tumorales de cáncer gástrico y de su rol en la expresión genética.

La expresión genética está estrechamente relacionada con la organización espacial del genoma

Las VE pueden afectar la organización de los TADs poniendo en contacto elementos de regulación genómica que normalmente no interactúan. Por ejemplo, la fusión de TADs adyacentes al perder o cambiar la posición relativa de sus secuencias frontera puede resultar en el contacto ectópico entre potenciadores y promotores de genes que normalmente no son sus blancos, derivando en la expresión desregulada de los mismos[7,18].

Muchos genes relevantes para el cáncer han sido estudiados en su contexto genómico y se conoce su ubicación en asas de cromatina interaccionando con sus elementos reguladores. Estos genes incluyen pero no se limitan a *KRAS*, *NRAS* y *BRAF*, que son miembros de la vía RAS y RAF; MYC, el oncogén humano más amplificado y sobreexpresado con mayor frecuencia; *TP53* que codifica para la proteína p53 que es el gen más frecuentemente mutado en los distintos tipos de

cáncer; *EGFR*, que codifica el receptor del factor de crecimiento epidérmico, un blanco relevante para distintos fármacos y el ligando de muerte programada PDL1 una proteína blanco para inmunoterapia[25], por mencionar algunos ejemplos para poner en contexto la importancia de la topología en la regulación genética (Figura 6).

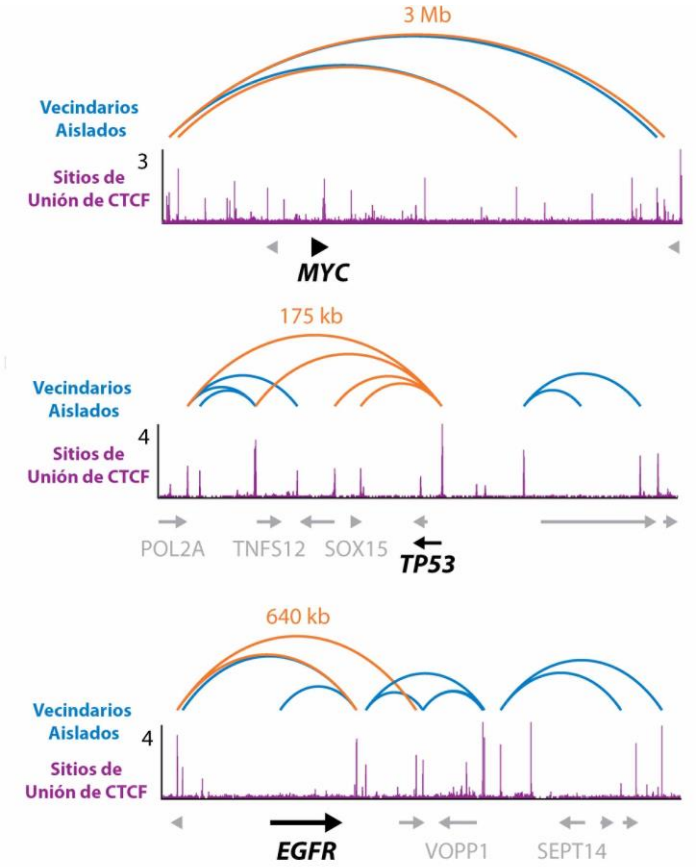


Figura 6. Ejemplos de genes relevantes en el cáncer en su contexto genético. Los arcos representan interacciones entre *loci* (azul) y las interacciones que presentan los genes de interés MYC, TP53 y EGFR con sus elementos reguladores en naranja. Cada ejemplo contiene un track de picos de CTCF (morado) marcando las fronteras de los vecindarios aislados o asas de cromatina.

Además, otro factor importante para la ubicación de CTCF en el material genético es la metilación de la cromatina proceso que se ve afectado durante la carcinogénesis, provocando un cambio en las fronteras delimitadas por CTCF ya que en general este elemento de frontera no se une a regiones metiladas y en consecuencia se generan nuevos contactos[25].

Por ejemplo, el caso de los gliomas asociados a mutaciones en genes de la isocitrato deshidrogenasa (IDH) en los que el nivel de metilación del ADN aumenta globalmente, los sitios CTCF ubicados en una región frontera de TAD cerca del receptor-α del factor de crecimiento derivado de plaquetas (PDGFRA) se metilan, lo que conduce a una disminución de la unión de CTCF a esta región y a la expresión de PDGFRA por un potenciador ubicado en el TAD adyacente. En el mismo estudio se depletaron las fronteras de CTCF mediante CRISPR obteniendo la misma interacción ectópica y para comprobar el mecanismo se desmetiló el genoma de estas células con 5'azacitidina, inhibiendo la expresión de este gen por la unión normal de CTCF a las fronteras del TAD que lo contiene[62]. Lo anterior muestra la plasticidad de estas interacciones y como pueden ser afectadas en el cáncer sin necesidad de mutación.

Además, en estudios recientes con líneas celulares de cáncer de páncreas se ha demostrado que la topología tiene un impacto importante en la habilidad de las células de entrar en metástasis, un marcador que afecta fuertemente la sobrevida de los pacientes en distintos tipos de cáncer[63]. Al realizar experimentos de Hi-C en una línea celular de cáncer de páncreas primario (PANC-1) y comparar su topología con una línea celular metastásica del mismo cáncer (Capan-1) encontraron nuevas interacciones entre potenciadores y promotores en estas últimas (Figura 7).

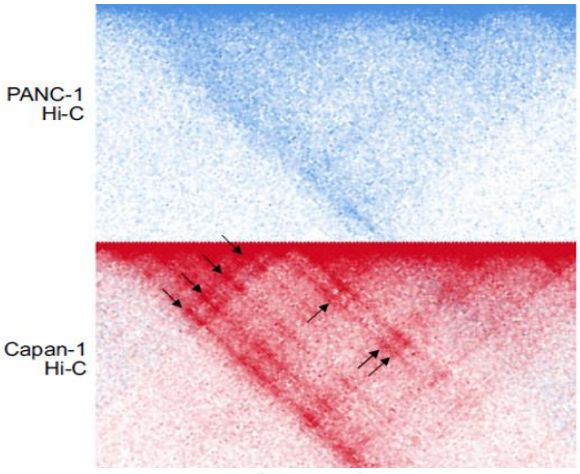


Figura 7. Matrices de Hi-C de una región genómica del cromosoma 15 de la líneas celulares de cáncer de páncreas PANC-1 y Capan-1. Corresponden a cáncer primario y secundario respectivamente, donde se pueden observar nuevos contactos en la línea metastásica (Modificada de Ren et al., 2021[63]).

En conclusión, las variantes estructurales (VE) pueden tener un impacto significativo en la organización de los Dominios Topológicamente Asociados (TADs), al poner en contacto elementos de regulación genómica que normalmente no interactúan. La fusión de TADs adyacentes, debido a cambios en las secuencias frontera, puede provocar un contacto ectópico entre potenciadores y promotores de genes, lo que conduce a una expresión desregulada de los mismos.

Impacto de las variaciones estructurales en el genoma del cáncer

Algunos estudios han demostrado que la fusión de los TADs y el mezclado de estos por VE son relevantes en la oncogénesis. En pacientes con leucemia mieloide aguda, la reubicación de un potenciador por inversión en el cromosoma 3 activa el protooncogén *EVI1* [18].

En cáncer gástrico, la duplicación en tándem de un fragmento de la región promotora del gen *CCNE1* que se enriquece con la marca de histonas H3K27ac y comienza a actuar como potenciador, lleva a la sobreexpresión de *CCNE1* que está relacionada clínicamente con una mayor tasa de mortalidad y resistencia a distintos fármacos en pacientes con CG. Además se asocio con la formación de TADs *de novo* e interacciones ectópicas en el tumor T2000877 como lo llamaron los autores Ooi et al., 2020. Esto se muestra en la figura 8 en un modelo (A) y con las matrices de Hi-C (B) donde esta VE fue detectada[7] comparando con la línea celular de CG SNU16.

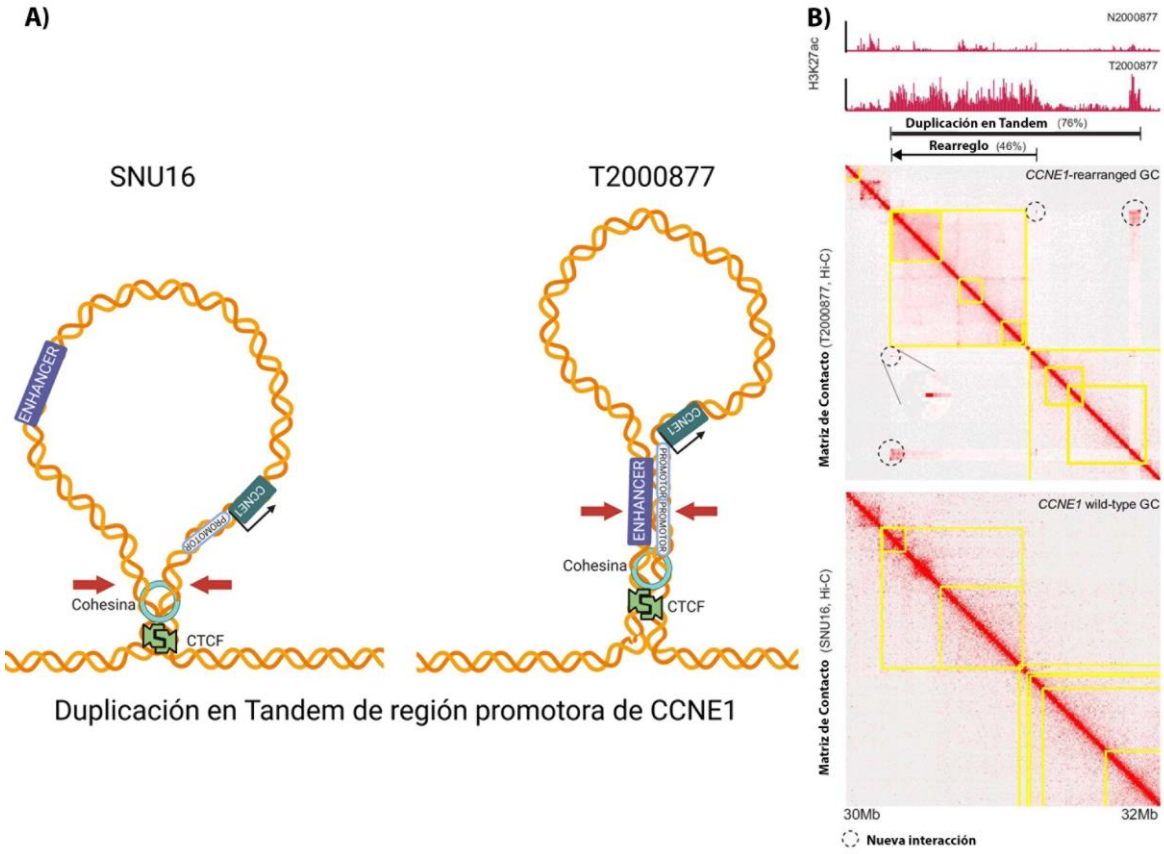


Figura 8. Modelo y matrices de la duplicación parcial en tándem de la región promotora del gen CCNE1. A) Modelo de las interacciones ectópicas provocadas por la duplicación parcial en tándem de la región promotora de *CCNE1*.**B)** Matrices de Hi-C de la región donde ocurre esta duplicación en el cromosoma 19, se muestran los TADs detectados en amarillo y las interacciones *de novo* marcadas con círculos punteados. Se compara el tumor de CG T2000877 con la línea celular de CG SNU16. Arriba se muestra el track de H3K27ac en la región mostrando un enriquecimiento en T2000877 (Modificada de Ooi et al., 2020[7], modelo realizado en Biorender).

En las matrices de Hi-C podemos detectar patrones característicos de diferentes VE ya que las interacciones ectópicas generan enriquecimientos de los contactos genómicos en trans cuya frecuencia no es esperada en un genoma que no ha sido rearreglado (Figura 9) y a través de la implementación de flujos de trabajo bioinformáticos específicos se puede llevar a cabo una caracterización sistemática de la VE presentes en una muestra a nivel de todo el genoma.

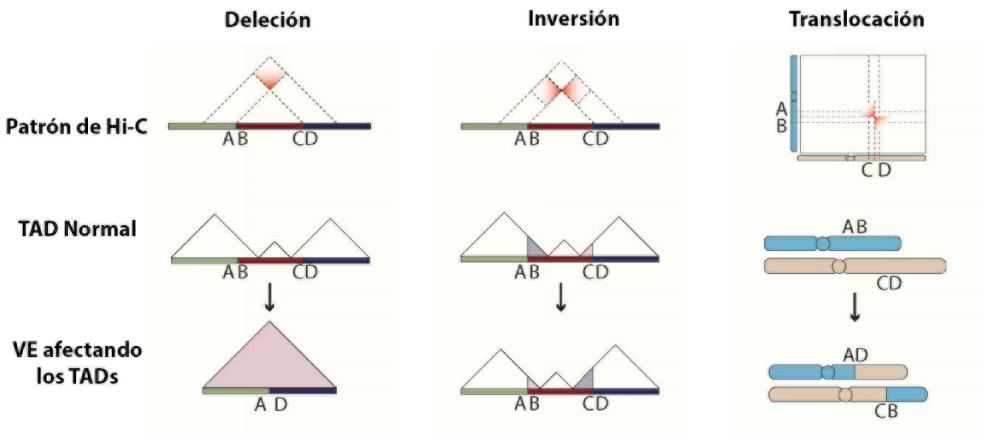


Figura 9. Ejemplos de patrones observados en las matrices de Hi-C para las VE. Se muestran los patrones de deleción, inversión y translocación en los datos de Hi-C y los modelos de los cambios producidos por las mismas a nivel de TADs y cromosomas (Modificada de Kim et al, 2019[18]).

Las características moleculares del cáncer son diversas, y la inestabilidad genómica es un factor que contribuye a esta diversidad. Es relevante destacar que la inestabilidad genómica tiene implicaciones en la supervivencia de pacientes con cáncer gástrico (CG) de tipo intestinal, como se muestra en los datos presentados por Ooi et al. en 2020 (Ver figura 8).

En algunos casos, pueden presentarse contactos ectópicos sin la presencia de variaciones estructurales (VE) o mutaciones, lo cual puede ser una respuesta a procesos cancerígenos como la metástasis o la proliferación celular. Esto subraya la complejidad de la regulación genética en el cáncer, lo que hace que el estudio de esta enfermedad mediante herramientas como el Hi-C sea invaluable. Esta técnica proporciona información global sobre la arquitectura del genoma, revelando, por ejemplo, la presencia de VE. Además, permite el estudio detallado de la regulación genética en estas patologías, demostrando así su versatilidad y su importancia para comprender la expresión de los genes.

Integración de datos topológicos (Hi-C) y transcripcionales (scRNA-seq)

El objetivo de este proyecto es estandarizar la técnica de Hi-C en muestras de tejido de cáncer gástrico del tipo difuso, así como en su tejido adyacente sano y células mononucleares de sangre periférica (PMBC) del mismo paciente. Esto permitirá identificar los cambios en la topología del genoma a nivel de TADs y VE. Además, se busca integrar datos de secuenciación de RNA de células individuales (scRNA-seq) derivados del tejido adyacente y del tumor de cáncer gástrico difuso del mismo paciente, enriqueciendo así las observaciones obtenidas.

En el scRNA-seq se utiliza un chip de microfluidos que contiene perlas de gel con secuencias de oligonucleótidos únicas. Estas perlas de gel están fusionadas con primers que se unen a los transcritos de ARN. Las células individuales se encapsulan en gotas de aceite junto con las perlas de gel, donde se produce la lisis celular y la liberación de ARN. Dentro de estas gotas, se forman pequeñas "gemas" que contienen una perla de gel, una célula y sus transcritos de ARN. Luego, se lleva a cabo la amplificación de ARN mediante la transcripción inversa y la amplificación de ADN complementario (cADN) utilizando primers específicos. Posteriormente, se realiza la secuenciación de alto rendimiento para obtener perfiles de expresión génica a nivel de célula única[64] y conocer las identidades celulares que conforman a los tejidos estudiados[65,66]. El protocolo para realizar esta técnica se describirá con detalle en la sección de materiales y métodos.

Cabe mencionar que este proyecto se lleva a cabo en colaboración con el Laboratorio de Biología Celular y del Desarrollo del Instituto de Fisiología Celular (IFC), con quienes se trabajó estrechamente para obtener las bibliotecas de scRNA-seq de estas muestras, este grupo está centrado en conocer las poblaciones inmunes presentes en el cáncer gástrico. Una vez secuenciadas, se realizó el análisis bioinformático correspondiente. El objetivo es caracterizar las relaciones entre la topología del genoma y la regulación genética en el cáncer gástrico difuso,

utilizando como herramienta los datos de secuenciación de las bibliotecas de Hi-C y las alteraciones transcripcionales detectadas por scRNA-seq.

Este enfoque permitirá proponer estrategias para la integración de estos datos genómicos, y continuar trabajando con más muestras de CG en el laboratorio en el futuro. Al detectar locus que presenten alteraciones topológicas y correlacionen con cambios en la expresión genética, se identificarán regiones candidatas de estudio en cáncer gástrico. Estos hallazgos contribuirán al conocimiento de los mecanismos moleculares que subyacen a la desregulación de la expresión de genes en esta patología, la cual es de gran relevancia en nuestro país.

Planteamiento del problema

El cáncer gástrico es un problema de salud pública relevante en nuestro país ya que representa la tercera causa de muerte por cáncer[11]. Además, este padecimiento se ha asociado con poblaciones hispanas [6,67], lo que podría estar relacionado con causas genéticas. En particular el cáncer gástrico de tipo difuso tiene una sobrevida a 5 años después del diagnóstico de la enfermedad que no supera el 10%[3] debido a la falta de tratamientos dirigidos y eficaces. La heterogeneidad genética del cáncer gástrico difuso está poco estudiada en el mundo y a la fecha no existen estudios que profundicen en la arquitectura genómica y la regulación transcripcional de esta patología en México. Caracterizar de manera precisa la topología y las alteraciones cromosómicas que afectan la organización del genoma y en consecuencia la expresión génica en este tipo de cáncer sienta las bases de conocimiento que pueden conducir al desarrollo de mejores herramientas de diagnóstico y tratamiento.

Hipótesis

El tumor de cáncer gástrico difuso tendrá alteraciones en la topología genómica que afectarán la expresión de genes, en contraste con el tejido adyacente.

Objetivo general

Identificar Dominios Topológicamente Asociados (TADs) alterados y Variaciones Estructurales (VE) en un tumor de cáncer gástrico del tipo difuso, comparándolo con su tejido adyacente y las células mononucleares de sangre periférica (PMBCs) del mismo paciente. Además, se busca correlacionar los datos topológicos con los datos de transcripción obtenidos mediante scRNA-seq del mismo tejido tumoral y tejido adyacente, para comenzar una caracterización detallada de la topología genómica y su impacto en la regulación genética en este tipo de cáncer.

Objetivos Particulares

1. Estandarización de la técnica de Hi-C en PMBCs, tejido adyacente y cáncer gástrico difuso del mismo paciente.

2. Implementación y estandarización de los flujos de trabajo bioinformático para identificar los TADs y VE en datos de Hi-C así como para el análisis de datos de scRNA-seq.

3. Comparación de datos de Hi-C de cáncer gástrico previamente publicados con los datos obtenidos para evaluar la eficiencia de los experimentos y la sensibilidad de detección de VE.

4. Integración de los datos obtenidos mediante Hi-C y scRNA-seq.

5. Identificar loci relevantes para el estudio del cáncer gástrico difuso tomando en cuenta los datos topológicos y transcripcionales obtenidos.

Materiales y Métodos

Obtención de Muestras

Las muestras de tumores, el tejido adyacente y la sangre periférica se obtuvieron de pacientes con cáncer gástrico que participaron en el proyecto “Identificación de subtipos celulares, vías moleculares y paisajes cromosómicos asociados con la progresión y malignidad tumoral” que se realizó en el Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ) mediante un protocolo aprobado por su comité de ética en investigación “HEM-3274”, en el que los pacientes dieron su consentimiento antes de ser incluidos en el presente estudio (Ver anexo de carta de consentimiento informado).

Los criterios que se siguieron incluyen a pacientes mayores de 18 años con sospecha o diagnóstico de cáncer gástrico epitelial, en cualquier etapa de cáncer invasor al diagnóstico (I a IV) que no hubieran recibido quimioterapia, radioterapia u hormonoterapia previamente y sin presencia de infección activa, enfermedad o tratamiento que cause inmunosupresión. Cabe mencionar que las muestras fueron obtenidas de dos maneras, mediante la realización de endoscopía diagnostica donde se adquirieron biopsias de los tejidos anteriormente mencionados o por la cirugía de resección del tumor y tejido adyacente.

Disgregación celular

Las muestras de los tejidos frescos provenientes de endoscopía o cirugía fueron transportadas en medio RPMI suplementado con 10% de SFB en hielo, aproximadamente una hora después de ser obtenidas al Instituto de Fisiología Celular donde fueron pesadas en seco, se eliminó el tejido que contenía sangre o grasa y fueron cortadas cuidadosamente en trozos cercanos a 1 mm, posteriormente fueron lavadas con PBS 1x y finalmente fueron tratadas con una solución rica en enzimas proteolíticas y colagenolíticas según reporta el fabricante,

Accumax (Merk) de 20 a 30min dependiendo de la cantidad de tejido para obtener suspensiones de célula única.

El número celular y viabilidad de estas fue evaluado mediante hematocitómetro con un volumen de 10µL de la disolución celular conteniendo azul de tripano y por citometría de flujo marcando las células con DAPI.

La búsqueda de alta viabilidad es relevante debido a que estas células fueron utilizadas tanto para el protocolo de Hi-C como el de scRNA-seq ya que este proyecto se lleva a cabo en colaboración con el Laboratorio de Biología Celular y del Desarrollo del IFC.

Una vez obtenida la suspensión de célula única se procedió a realizar el ensayo de Hi-C.

Captura conformacional cromosomal de alto rendimiento, “Hi-C”

Los volúmenes utilizados en este protocolo consideran un experimento con ~1 millón de células y se utilizaron tubos LoBind debido al número celular.

Fijación de la cromatina

Las células se fijaron con 2% de formaldehído por 10 min en agitación y se frenó la reacción con glicina 1M. Después de dos lavados con PBS 1X en hielo las células fijadas son centrifugadas a 300 x g por 10 min a 4°C, se desechó el sobrenadante, fueron congeladas con nitrógeno líquido y se almacenaron a -80˚C.

Permeabilización y obtención de núcleos

Las células se permeabilizaron con un buffer hipotónico (Tris-HCL pH 8 10 mM, IGEPAL 0.2%, NaCl 5M, pepstatina 1000X, inhibidor de proteasas fluoruro de fenil-metil-sulfonilo (PMSF) 100X y buffer litio-acetato-borato (LAB) 100X) en un volumen de 500µL durante 30min, fueron centrifugadas a 300 x g por 10min a 4°C y se resuspendieron en 500µL de buffer NEB2B 1X, se centrifugaron nuevamente a 300 x g por 10min a 4°C, se descartó el sobrenadante y se agregaron 50µL de una

solución de NEB2 1X y 0.3% de SDS, se incubaron durante 45min a 37°C sin agitación. Posteriormente, se agregaron 170µL de Triton X-100 al 1.14% y se tomaron 12µL para el control “no digerido”.

Digestión de la cromatina

La cromatina fue digerida con la enzima DpnII cuyo motivo de restricción es 5´…GATC…3´ 3´…CTAG…5´, en buffer 20µL de NEB2.1 10X y se consideraron 400U de la enzima que fueron agregados en dos momentos. Primero se dejaron 200U incubando toda la noche a 37°C y al siguiente día fueron agregadas otras 200U que se incubaron bajo las misma condiciones por 4h. Posteriormente, la enzima fue inactivada llevando los tubos a 62°C durante 20min. En esta etapa se tomó una alícuota de 12µL para el control “digerido”.

Biotinilación y reparación

Los fragmentos de ADN digeridos fueron reparados con 18.75µL de biotin-14-dATP a 0.4 mM, 0.75µL de cada nucleótido restante a una concentración de 0.4 mM y 8µL de DNA Polimerasa 1 subunidad Klenow 5U/µL. La reacción fue incubada durante 75 min a 37°C con 300 rpm de agitación.

Ligación por proximidad

Los fragmentos obtenidos fueron ligados *in nuclei* por la ADN ligasa T4. Se generó un master mix de 768µL de agua libre de nucleasas, 150µL de NEB T4-DNA-buffer de ligasa 10X, 120µL de Tritón X-100 y 36µL BSA, el cual fue agregado a las muestras en un volumen de 748µL junto con 10µL de ADN ligasa T4 (5 Weiss U/µL). La reacción se incubó toda la noche a 20°C.

Reversión del Crosslink

Para eliminar las proteínas y el ARN andamio que mantiene las interacciones entre la cromatina y fueron fijados por el formaldehído, se agregaron 40µL de proteinasa

K (10mg/mL) y 2µL de ARNasa I (10mg/mL) y se incubaron a 37°C por 2h, a 55°C por 1 h y se dejaron a 65°C sin agitación toda la noche.

Extracción del ADN por fenol/cloroformo

Se extrajo el ADN agregando 500µL de un mix de fenol-cloroformo-isoamyl alcohol (alcohol PCI) (25:24:1) y se agitó hasta que fue obtenida una fase blanca homogénea, se centrifugaron las muestras a 1500rpm por 10min a temperatura ambiente y se separó en otro tubo la fase acuosa, este proceso se repitió agregando 100µL de buffer Tris-HCL 1 M EDTA 0.5 M pH 8 (TLE).

Posteriormente, se agregaron 55µL de Acetato de sodio 3M pH 5.2, 4µL de glicógeno 20mg/mL y 2X de etanol puro frío. Se mezcló por inversión y fue incubado por 15min a -80°C. Se dejaron las muestras toda la noche a -20°C.

Al siguiente día se centrifugaron las muestras a 15000 rpm a 4°C por 30min, se descartó el sobrenadante y se lavó el pellet con etanol frío al 70% dos veces, centrifugando a 15000 rpm a 4°C por 10min. Fue removido el etanol y se dejó secar el pellet, después se disolvió en 20µL de TLE (10 mM Tris-HCl, 0.1 mM EDTA) y fué incubado por 5min a 37°C. La concentración de ADN fue cuantificada usando un Qubit, que calcula la concentración de ADN por fluorescencia.

Sonicación de las moléculas circulares

Las muestras se llevaron a un volumen de 130µL de TLE y se utilizó un sonicador Covaris para obtener moléculas lineales durante 55 s a 200 ciclos por ráfaga y una amplitud (w) de 140.

Pull-Down de Biotina

Se prepararon los siguientes buffers:

2X NTB (No Tween Buffer) (10 mM de Tris-HCl pH 8, 1mM EDTA ,1M NaCl), 1X NTB (concentraciones a la mitad).

1XTB (Tween Buffer) (5mM de Tris-HCI pH 8, 0.5 mM EDTA, 1M NaCI, 0.05% Tween20), 0.5X TB (concentraciones a la mitad).

Se tomaron 100µL de perlas Dynabeads MyOne Streptavidin C1 (Invitrogen) por muestra. Las muestras fueron lavadas 2 veces con 1mL de 1X TB mezclándolas por rotación a temperatura ambiente por 2min. Fueron resuspendidas en 400µL de 2X NTB, se agregaron al ADN de Hi-C y se incubaron por 30min a temperatura ambiente en rotación.

Las perlas fueron separadas del sobrenadante mediante una gradilla magnética y fueron realizados 2 lavados con 400µL de 1X NTB en rotación a temperatura ambiente por 3 min. Las perlas fueron lavadas con 400µL de 0.5X TB, se incubaron a 55°C por 3min con agitación a 750rpm. Finalmente se lavaron con 100µL de 1X buffer de ligación (NEB) por rotación durante 3min.

Remoción de Biotina

Las perlas fueron resuspendidas en 6µL de Buffer de ligación 10X NEB, 2µL de 10 mM dATP y 5µL de T4 DNA polimerasa. Se incubó la reacción a 20°C por 30min sin rotación.

Reparación de extremos

Se agregaron 5µL de 10 mM de dNTPs, 6 µL de Buffer de ligación10X NEB, 5µL de T4 PNK (10U/µL) y 1µL de Klenow, fueron incubadas las muestras por 30min a 20°C, se lavaron las perlas con 400µL de 0.5X TB incubándose a 55°C por 3 minutos con 750rpm de agitación. Por último, las perlas fueron lavadas con 100µL de 1X NEB2 por rotación durante 3min.

Posteriormente, a las muestras se les agregaron 5µL de dATP, 10µL de 10X NEB2 y 5µL de Klenow exo^{-}, se incubaron a 37°C por 30min sin rotación, Fue removido el sobrenadante y las perlas fueron lavadas 2 veces con 400µL de 0.5 TB incubándose a 55°C por 3min con agitación a 750rpm. Se lavaron las perlas con 400µL de 1X NTB en rotación a temperatura ambiente por 3min, fueron lavadas con 100µL de

Buffer de ligación 1X por rotación a temperatura ambiente por 3min. Finalmente, se resuspendieron las perlas en 50µL de Buffer de ligación 1X.

Preparación de biblioteca

Se agregaron 4µL de adaptadores Truseq (15mM) y 2.4µL de ADN ligasa T4 (5U Weiss/µL). Finalmente, el éxito de esta etapa es probado amplificando las bibliotecas con primers para los adaptadores, se seleccionaron los tamaños adecuados para la secuenciación mediante perlas SPRI (solid phase reversible immobilization beads) y fueron comprobados por High Sensitivity D1000 ScreenTape (Agilent). Finalmente, las bibliotecas se mandaron a secuenciar en un equipo Novaseq de ilumina para obtener un aproximado de 600 millones de lecturas.

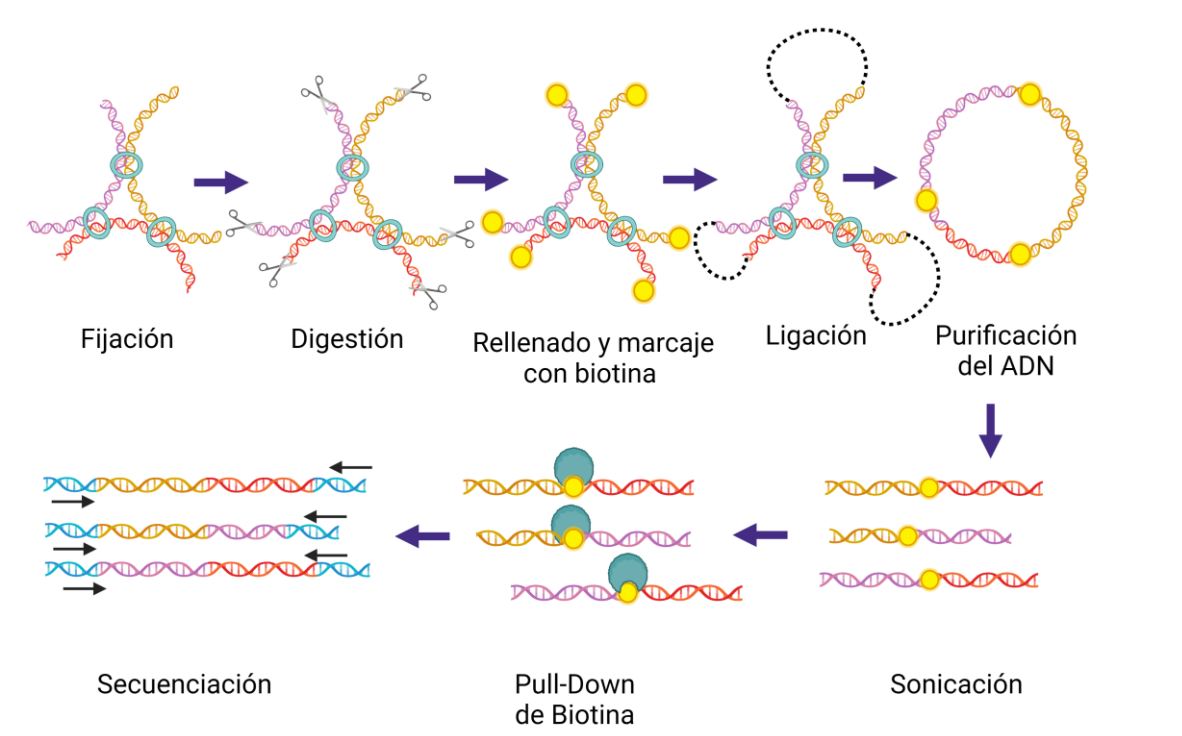


Figura 10. Diagrama general de las etapas principales del protocolo de Hi-C. Modelo de las formas de las moléculas involucradas (figura realizada en Biorender).

Cabe destacar que el protocolo general de la técnica debe ser acoplado en los volúmenes utilizados dependiendo del número celular obtenido de los tejidos, que regularmente es bajo para las muestras de tejidos obtenidos por biopsias y se debe buscar el mayor desempeño para la obtención de material genético para Hi-C.

Cultivo de las Células AGS

La línea celular AGS de adenocarcinoma gástrico con referencia de ATCC CRL-1739, se mantuvo en cultivo con medio DMEM suplementado con 8% de suero fetal bovino (Biowest), 2mM de L-glutamina (Biowest), 100 U/mL de penicilina-estreptomicina (Biowest), y 2mM de piruvato de sodio (Corning). Los cultivos se incubaron a 37°C, 5% CO_2. El mantenimiento se realizó cambiando de medio cada 2 días y lavando con PBS 1x antes de colocar medio nuevo.

Cultivo de células K562

La Línea celular no adherente K562 de leucemia mieloide con referencia de ATCC CRL-3344, se mantuvo en cultivo con medio de Dulbecco modificado de Isocove (IMDM). Los cultivos se incubaron a 37°C, 5% CO_2. El mantenimiento se realizó cambiando de medio cada 2 días desechando el medio anterior y colocando nuevo.

Metodología experimental scRNA-seq.

Las suspensiones de células únicas de las muestras fueron procesadas mediante el protocolo de aislamiento basado en Droplets 10XChromium[64] que utiliza un chip de microfluídos que encapsula (Figura 11):

Perlas individuales de gel (figura 11 A) que contienen oligonucleótidos conformados por las siguientes unidades: 1) adaptadores (TruSeq) y primers para la secuenciación, 2) una secuencia barcode de 14pb que identificará a todos los transcritos pertenecientes a una célula, 3) Identificadores moleculares únicos (UMIS) que son secuencias de 10pb cuyo propósito es eliminar el ruido técnico producido por las dúplicas de PCR y poder identificar sin sobreestimar los transcritos presentes en la células. Es decir cada perla de gel individual contiene un solo

barcode pero este a su vez está unido a una amplia gama de moléculas UMIS (Figura 11 A). 4) un Poly(dT) de 30pb que se utilizará como primer para sintetizar el cADN de los transcritos de ARN poliadenilado. Por otro lado, la células en suspensión, las enzimas y reactivos que llevan a cabo la reacción de retrotranscripción son cargados en otro pozo del chip (Figura 11 B[64]).

Posteriormente, el glicerol que es cargado en otro pozo es empujado a través de los micro canales por el equipo ChromiumX formando "Gemas" que contienen individualmente una célula, una perla de gel y todos los reactivos para llevar a cabo la retrotanscripción dentro de la Gema. En esta etapa una emulsión parcial afecta directamente al número de células y lecturas únicas que serán obtenidas del experimento (Figura 11 B).

Finalmente, se rompe la emulsión, se amplifica el cADN y se construye la librería. Dando como resultado moléculas que contienen adaptadores de secuenciación, el barcode de la célula a la que pertenece el transcrito, UMI que identifica que tal molécula no pertenece a una sobreestimación de duplicados de PCR, el inserto de cADN y el índice de la muestra a la que pertenece. En una longitud adecuada para la secuenciación por Ilumina (Figura 11 C).

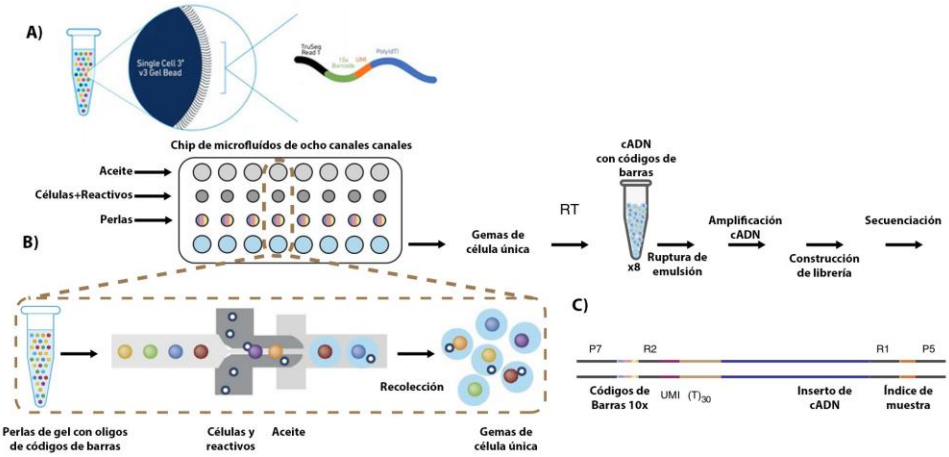


Figura 11. Flujo de trabajo en el experimento de scRNA-seq. A) Perlas de gel que contienen adaptadores TruSeq, barcode único por perla, UMI (Identificadores moleculares únicos) que permite distinguir del ruido técnico causado por las dúplicas de PCR, y Poly(dT) que será utilizada como primer para la reacción de rtPCR. **B)** Chip de microfluidos de Chromium 10X mostrando pozos donde se carga el glicerol (aceite), las células y las perlas de gel. Así como el flujo de trabajo de la generación de librerías. **C)** Moléculas resultantes del procesamiento mediante esta técnica, los elementos que las componen y que las hacen aptas para la secuenciación en plataformas ilumina (Modificada de Zheng *et al*, 2017[64]).

Análisis Bioinformático

Todos los fastq fueron alineados al genoma de referencia hg38. Los datos de Ooi *et al* 2020 de las librerías de Hi-C SNU16 (línea celular de CG) y T2000877 (tumor de CG intestinal) fueron obtenidos de la base de datos del Laboratorio Europeo de Biología Molecular-Instituto Europeo de Bioinformática (EMBL-EBI por sus siglas en inglés) en el repositorio con número PRJNA485437, para la comparación de calidad con los datos de Hi-C obtenidos en este trabajo.
Calidad de experimentos de Hi-C.

HiCUP

Se utilizó el flujo de trabajo bioinformático "pipeline" HiCUP[68] para obtener la calidad de las librerías de Hi-C analizadas ya que este detecta los artefactos inherentes al experimento y la estandarización del protocolo es uno de los objetivos de este trabajo.

Primero identifica los reads que corresponden a moléculas de Hi-C, ya que estas contienen un sitio de restricción generado por la digestión y ligación, mediante un módulo del pipeline llamado hicup_truncater. Posteriormente, se mapean estos reads al genoma de referencia utilizando índices de BOWTIE2[69], en el módulo hicup_mapper. Después se genera una digestión *in silico* del genoma de referencia la cual se compara con los datos obtenidos en el experimento. Esto se lleva a cabo en el módulo hicup_filter. Finalmente, se identifican los artefactos en el módulo hicup_deduplicator (Figura 12).

A partir de estos datos se pueden obtener el número de unique-ditags y relacionar los artefactos con etapas del protocolo, detectando donde se puede mejorar el mismo y que tan eficiente fué el experimento mediante un reporte de control de calidad.

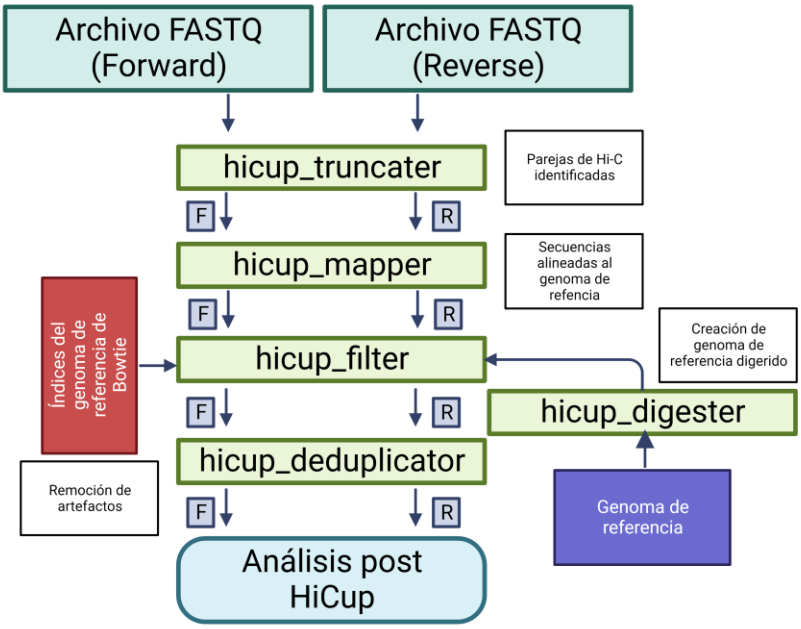


Figura 12. Diagrama de flujo de HiCUP. Se muestran los archivos de entrada y de salida requeridos para la alineación al genoma de referencia de las secuencias de Hi-C. Además de la digestión *in silico* requerida para la identificación de artefactos (Modificada de Wingett *et al*. 2015 [68]).

Generación de matrices de Hi-C, detección de CNV y translocaciones con HiNT[70]

HiNT[70] es un pipeline que se compone de tres módulos*: HiNT-Pre*, *HiNT-CNV* y *HiNT-TL*. Al analizar los datos de Hi-C con este se obtiene la matriz de contactos, la variación del número de copias y las translocaciones, respectivamente. A continuación se describirá cada módulo considerando que cada uno es un comando del pipeline (Ver figura 13 Diagrama de flujo del pipeline).

HiNT-Pre realiza el pre-procesamiento de los datos de Hi-C llevando a cabo el alineamiento de los pares de lecturas con BWA-MEM[71] (Figura 13) Siguiendo con

la generación de la matriz de contactos, almacenando las frecuencias de contacto entre dos *loci* genómicos tomando en cuenta las siguientes lecturas:

Pares de **lecturas normales de Hi-C,** las cuales son lecturas pareadas de fragmentos de ADN que alinean para dos *loci* distintos y están separados por un sitio de restricción modificado, siendo equivalentes con las unique Di-Tags antes mencionadas (Figura 14 diagrama 1).

Pares de **lecturas quiméricas no ambiguas** o unambiguous chimeric pairs en inglés. Estas corresponden a moléculas ligadas por Hi-C donde la lectura quimérica se compone de fragmentos A y B separados por el sitio de restricción. Sin embargo, el otro par de lectura alinea solo con el *locus* B (Figura 14 diagrama 2).

Finalmente, las **lecturas quiméricas ambiguas** que son también producto de ligación entre dos fragmentos genómicos que interaccionan, pero que existen alineando para un *locus* A y C antes del sitio de restricción modificado con un *locus* B como se muestra en el diagrama 3 de la figura 14. Por lo tanto, son útiles para la detección del breakpoint de las translocaciones.

La matriz resultante es un archivo .hic cuya característica principal es que este formato contiene múltiples resoluciones, facilitando su visualización mediante Juicebox[72] un programa dedicado a estas tareas. Todas las matrices mostradas en la sección de detección de VE fueron generadas y visualizadas de esta manera.

HiNT-CNV usa la matriz de contactos para predecir el número de copias de los segmentos genómicos. Brevemente, crea un perfil de cobertura a lo largo del genoma calculando las sumas de filas o columnas de la matriz de contacto a una resolución fija, por ejemplo, 50 kb. Estas sumas deben estar correlacionadas con el número de copias de los contactos ya que corresponden a la fuerza de interacción de esa región con todas las demás regiones.

Utiliza la matriz de contactos no normalizada, porque la normalización de equilibrio matricial (estableciendo la suma de cada fila o columna con un valor de 1) que es el enfoque más utilizado de normalización en Hi-C, no solo elimina los sesgos inherentes al experimento como el contenido de GC, mapeabilidad y la longitud de

los sitios de restricción, si no también la información del número de copias. Por lo tanto primero se genera el perfil de cobertura y luego se eliminan los sesgos.

Finalmente, por medio del algoritmo de segmentación BIC-seq2[73] identifica las lecturas enriquecidas o empobrecidas determinando las coordenadas de las variaciones del número de copias CNV.

Esto es representado en diagramas como el que se puede observar en la sección correspondiente a HiNT-CNV de la figura 14 donde se muestra cada cromosoma, considerando la cobertura de las lecturas presentes en la librería de Hi-C (Log2CopyRatio) en los intervalos log2 >= 0.3 en rojo y log2 <=-0.3 en verde son cada uno ganancia o pérdida de número de copias respectivamente. También en verde cercano a la línea cero en el intervalo entre el Log2CopyRatio (-0.3,0.3) se considera que no existe variación en el número de copias (Figura 14 sección HiNT-CNV).

Por último, el largo horizontal de cada línea representa a escala la longitud en pares de bases de la variación mostrando por ejemplo, en la sección HiNT-CNV de la figura 14 el cromosoma X duplicado.

HiNT-TL detecta las translocaciones analizando las interacciones intercromosómicas normalizadas de las matrices. En general, las probabilidades de contacto entre dos regiones en el mismo cromosoma disminuyen con la distancia y las interacciones intercromosómicas son menos frecuentes en comparación con las intracromosómicas. Cuando una translocación intercromosómica ocurre, se observa un aumento de probabilidad de contacto en los cuadrantes correspondientes a las interacciones entre dos cromosomas (Figura 14 sección HiNT-TL). Para identificar los breakpoints exactos, HiNT-TL primero busca las regiones translocadas con una resolución gruesa de 100kbp y refina a un solo par de bases utilizando las lecturas quiméricas ambiguas.

https://github.com/parklab/HiNT

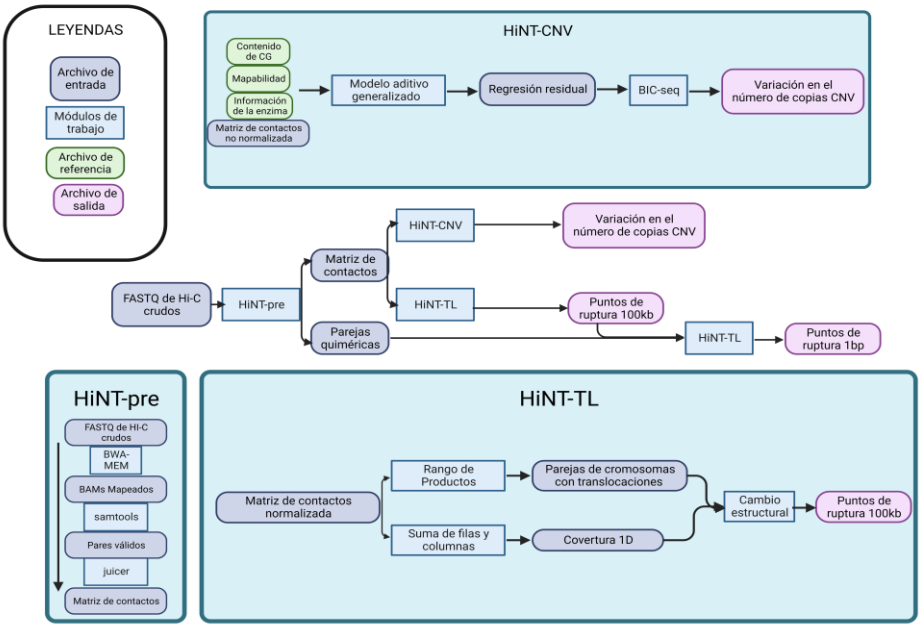


Figura 13. Diagrama de flujo de HiNT. Se representan los archivos de entrada y de salida, además de los módulos que requiere el pipeline para alinear el genoma, generar la matriz de contacto a partir de los datos de HiC, calcular el número de copias y el breakpoint a 100kb y 1bp (Modificada de Wang et al. 2020[70]).

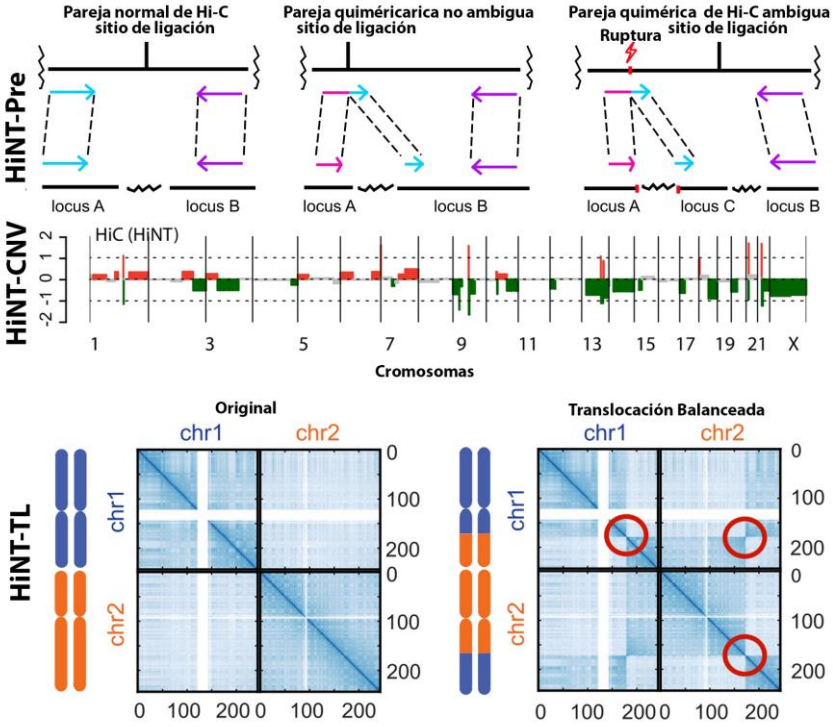


Figura 14. Módulos de HiNT. HiNT Pre genera la matriz de contactos utilizando las lecturas de Hi-C normales, lecturas quiméricas no ambiguas y las quiméricas ambiguas. **HiNT-CNV** calcula el número de copias partiendo de la cobertura y **HiNT-TL** ubica las translocaciones primero a una resolución gruesa de 100kb y utilizando los pares de lecturas quiméricas ambiguas para determinar el breakpoint de la translocación a la resolución de un solo par de bases (Modificada de Wang et al. 2020[70]).

HiCexplorer[74,75]

Las librerías de Hi-C SNU16, T2000877 y PMBC, tejido adyacente y tumor del paciente 8 fueron analizadas con el pipeline HiCexplorer[74,75] que contiene una amplia gama de herramientas informáticas que permiten la generación de matrices, la comparación entre bibliotecas, la identificación y visualización de TADs.

Se alinean los datos de Hi-C al genoma de referencia con Bowtie2[69] una vez obtenidos los archivos .bam se generan las matrices considerando el sitio de restricción mediante la herramienta *hicBuildMatrix* (Las matrices fueron corregidas con la opción default de HiCexplorer[74,75], False Discovery Rate "FDR" que detecta y descarta errores tipo 1 y requiere un valor q) y se normaliza el número de lecturas para cada matriz respecto a los datos con el número más bajo de lecturas para no sobreestimar los contactos en los análisis posteriores al comparar entre matrices, esto se realiza con el comando *hicNormalize*.

Posteriormente, se corrigen las matrices respecto al número de lecturas consideradas por contacto de Hi-C de tal manera que no se tomen en cuenta contactos de baja cobertura con *hicCorrectMatrix*. Una vez realizadas estas normalizaciones se procedió a llamar TADs mediante *hicFindTADs* utilizando como métrica el separation score o valor de separación para asignar fronteras entre TADs implementado en HiCexplorer[74,75] este se basa en el uso del promedio del z score de toda la matriz y define las fronteras entre TADs cuando este valor es alto y las regiones interiores de los TADs las que tienen un puntaje mínimo de separation score[74,75].

Considerando el número de unique-ditags los TADs fueron llamados a una resolución de 50kb. Por último se utilizó *hicDifferentialTAD* para encontrar los TAD´s diferenciales entre el tejido adyacente sano y el tumor del paciente 8, esta herramienta fue utilizada con los parámetros default del pipeline como se muestra en el ejemplo de uso en su repositorio, con una p=0.05 y un modo de rechazo de hipótesis nula (mode reject) en la opción "one" es decir, se consideran TADs diferenciales si al menos una región dentro del mismo tiene una p <= 0.05.

https://hicexplorer.readthedocs.io/en/latest/content/list-of-tools.html.

Análisis Bioinformático de los datos de scRNA-seq.

10XCellRanger

Se analizaron los datos del tejido sano adyacente y el tumor gástrico difuso del paciente 8 con el pipeline de 10XCellRanger que realiza mediante el comando *cellranger count*, el alineamiento y una descripción de la calidad de los datos, además arroja listas de genes detectados por cluster los cuales define mediante tSNE[76] que es un algoritmo que pretende mostrar gráficamente las relaciones que guardan los datos multidimensionales de scRNA-seq (esto se describirá con detalle posteriormente) y los archivos: **Features** que contiene la lista de los genes que se encontraron en las células del experimento, **Barcodes** que permite etiquetar los transcritos pertenecientes a cada célula y por último el archivo **Matrix** que contiene el número de lecturas para cada uno de los transcritos detectados.

https://support.10xgenomics.com/single-cell-gene expression/software/pipelines/latest/what-is-cell-ranger

Seurat[77]

Tomando estos archivos se puede continuar con el análisis de scRNA-seq con la librería de R Seurat[77] que permite la detección mediante el comando *PercentageFeatureSet* y selección de células únicas y vivas respecto al porcentaje de expresión de genes mitocondriales con el comando *subset,* eliminando los datos que contienen menos de 200 genes detectados y manteniendo los que contienen más de 2500 genes con un porcentaje menor al 5% de expresión de genes mitocondriales, estos parámetros son estándar para estos datos.

Sin embargo, en el caso de la muestra de origen tumoral se amplió el porcentaje de genes mitocondriales al 20%[65,66,78,79] mientras que en el tejido sano se mantuvo en 5% (Ver figura 35). Estos datos se visualizaron en diagramas de violín de los genes detectados, el número de lecturas para cada gen y el porcentaje de genes mitocondriales mediante el comando *Vlnplot*.

Posteriormente se normalizaron los datos para cada biblioteca con *NormalizeData* logarítmicamente, se detectan los genes ampliamente variables con *FindVariableFeatures* (Ver figura 36) y se escalan mediante *escaleData* para hacer comparables los datos de expresión entre sí.

La multidimensionalidad en los experimentos de scRNA-seq

Siguiendo con las herramientas de Seurat y una vez normalizados los datos respecto a su varianza se procedió con la reducción de dimensiones.

El experimento de scRNA-seq permite conocer todos los transcritos expresados a nivel de célula única en miles de células[64]. Esto genera datos multidimensionales lo cual supone un reto estadístico, ya que se vuelve difícil distinguir las diferencias de expresión que definen a las poblaciones celulares de las diferencias inherentes a las mismas, explicadas por variación intercelular. Por lo tanto la reducción de dimensionalidad es esencial para el análisis de datos de scRNA-seq, permitiendo la representación gráfica en dos dimensiones de las relaciones que guardan las células entre sí, respecto a su perfil de expresión.

El análisis de componentes principales PCA es una herramienta de reducción de dimensiones lineal que crea nuevas variables “componentes” que buscan capturar la mayor cantidad de variación (información biológica) en los datos, dibujando las relaciones globales presentes en los mismos. Esto se realizó con los comandos de Seurat *RunPCA, FindNeighbors y FindClusters*.

Posteriormente, estos componentes pueden ser utilizados por otros algoritmos de reducción de dimensiones más eficientes para la captura de relaciones locales como t-SNE (T-distributed Stochastic Neighbor Embedding) y (Uniform manifold approximation and Project). Estos se diferencian en que t-SNE es estocástico, es decir, que las relaciones que recupera cada vez que es llevado a cabo pueden verse afectadas en un análisis posterior, además no conserva las relaciones globales de manera eficiente. Es por esto que se decidió realizar la reducción de dimensiones mediante UMAP (comando *RunUMAP*) que no es estocástico,

conserva las relaciones globales correctamente y además es computacionalmente más eficiente que t-SNE.

En concreto, t-SNE es la herramienta utilizada por el pipeline 10XCellRanger y el paquete de R Seurat[77] permite la utilización de UMAP. Se anotaron las identidades celulares de los clusters definidos por UMAP con la función *SingleR* utilizando la base de datos HeOrganAtlasData[80] que contiene datos de scRNA-seq de 84 mil células provenientes de 15 tejidos distintos de un solo donante, con el fin de llamar identidades celulares mediante las firmas transcripcionales de cada tipo celular presente en los tejidos adyacente y tumoral del paciente 8.

Detección de genes diferencialmente expresados mediante MAST

Finalmente, se utilizó mediante Seurat al algoritmo MAST[81] (Model-based analysis of single-cell transcriptomics) optimizado para la detección de genes diferencialmente expresados (GDE) entre muestras ya que recupera la información de la expresión de los transcritos sin subestimarla con el comando *FindMarkers*.

Esto es un problema que tienen otros algoritmos que consideran tomar los datos de scRNA-seq en bulk, ya que la expresión de transcritos entre poblaciones celulares es distinta, al analizar los perfiles de expresión la información para estos transcritos puede ser cero en un grupo de células y tener un nivel de expresión en otro grupo, el exceso de ceros puede llevar a la expresión de ciertos transcritos biológicamente relevantes a un nivel negativo, por lo tanto se perdería esta información. MAST[81] lo resuelve manteniendo la expresión de todos los transcritos por arriba del negativo, buscando no perder información biológica.

https://satijalab.org/seurat/

Análisis de enriquecimiento

Una vez obtenidas las listas de GDE con el algoritmo MAST[81] como se describe en la sección anterior, se utilizó la herramienta web Enrichr[82–84] para el análisis de enriquecimiento, que contiene 180,184 genes anotados provenientes de 102 librerías de genes. Además, permite la generación de una amplia gama de

ontologías, desde factores de transcripción y enfermedades hasta relación de los GDE con procesos biológicos.

En particular en el presente trabajo se manejó la ontología de procesos biológicos que utiliza la base de datos REACTOME[85,86] que tiene 14,516 interacciones entre proteínas anotadas, donde se consideraron los procesos biológicos más relevantes en los que están involucrados los genes sobreexpresados respecto a su score combinado, que se define por la siguiente ecuación: C= log(p)*z donde C es el score combinado, p= al valor de p de la prueba exacta de Fisher y z= al valor z para la desviación del rango esperado[82–84]. En términos biológicos considera la importancia de los genes en un proceso respecto a la base de datos y el valor de p ajustada que define su significancia.

https://maayanlab.cloud/Enrichr/

Integración de datos topológicos y transcripcionales

Para la integración de los datos topológicos y transcripcionales se utilizó el visualizador de datos genómicos Seqmonk, donde se agregaron las coordenadas de los genes diferencialmente expresados y de los TADs exclusivos del tumor utilizando las ubicaciones genómicas de cada dato como sondas (probes) se pudieron realizar listas de TADs diferenciales que contenían GDE y el número de GDEs contenidos en TADs diferenciales.

https://www.bioinformatics.babraham.ac.uk/training.html#advancedseqmonk

Las matrices de los locus que contienen los genes de las subunidades de la cohesina y sus reguladores, fueron ploteadas con la herramienta de HiCexplorer[74,75] *hicPlotTADs* considerando los TADs llamados previamente a 50kb de resolución en las matrices normalizadas y corregidas como se describe previamente, incluyendo las coordenadas de todos los genes presentes en los *locus*.

Resultados

Conjunto de muestras obtenidas y procesamiento inicial

Las muestras recibidas en el IFC son derivadas de biopsias obtenidas por endoscopía y en algunas ocasiones de cirugías de resección, estas son trasladadas al instituto en medio RPMI con 10% de SFB colocando las muestras en hielo. Posteriormente fueron procesadas con Accumax obteniendo disoluciones de célula única, fueron contadas y calculada su viabilidad mediante la tinción con azul de tripano en hematocitómetro y por citometría de flujo utilizando un marcaje con DAPI como se detalla en el apartado de materiales y métodos. Es importante recalcar que el número celular obtenido es dependiente del tamaño y peso de la biopsia o resección.

Debido a que la endoscopía es un protocolo de diagnóstico, también se han recibido muestras distintas al cáncer gástrico ya que estas deben ser procesadas de inmediato para mantener la viabilidad y el diagnóstico suele ser posterior, una vez realizado por los patólogos del NCMNSZ se determina si se continúa con el protocolo. Las muestras 18 y 19 corresponden a un pólipo gástrico y a una lesión gástrica, ejemplificando este punto (Ver tabla 1 con los datos de las muestras procesadas).

Cabe destacar que no siempre es posible obtener todos los tejidos por condiciones inherentes al tratamiento y salud de los pacientes. Sin embargo, estas biopsias serán utilizadas como controles (Ver tabla 1 paciente 004).

Tabla 1. Muestras de Cáncer gástrico fijadas para Hi-C con subtipo, número celular y viabilidad calculada con hematocitómetro y citometría.

	Diagnóstico	Peso de la Muestra	Número Celular	Viabilidad	# De células Fijadas para Hi-C
001	Adenocarcinoma gástrico intestinal/ Tipo tubular bien diferenciado grado I	Tumor: 0.04g Tejido Sano: 0.06g	Tumor: 1,880,000 Tejido Sano: 812,000	Tumor: 72% Tejido Sano: 85%	Tumor: ~ 1,000,000 Tejido Sano: ~500,000
002	Adenocarcinoma Difuso poco diferenciado con células de anillo de sello	Tumor: 0.7g Tejido Sano:0.23g	Tumor: 10,071,000 Tejido Sano: 1,230,000	Tumor: 93% Tejido Sano: 70%	Tumor: ~5,035,500 Tejido Sano: ~615,000
004	Adenocarcinoma de la unión esofagogástrica. Poco diferenciado	Tejido Sano: 0.04g	Tejido sano: 1,248,333	Tejido Sano: 68%	Tejido Sano: Aún no fijado.
008	Adenocarcinoma pobremente diferenciado del tipo difuso	Tumor: 1.01g Tejido Sano: 0.27g	Tumor: 2,060,000 Tejido Sano: 3,700,000	Tumor: 80% Sano: 80%	Tumor: 1,900,000 Tejido Sano: 3,200,000
015	Adenocarcinoma gástrico	Tumor: 0.01g Tejido Sano: 0.02g	Tumor: 276,450 Tejido Sano: 556,000	Tumor:72.5% Sano:86%	Tumor: 176,000 Tejido Sano: 400,000
018	Pólipo gástrico	Pólipo: 0.03g Tejido Sano: 0.02g	Pólipo: 1,210,000 Tejido Sano: 940,000	Pólipo:68% Sano:76%	Pólipo: 700,000 Tejido Sano: 500,000
019	Lesión gástrica no maligna	Lesión: 0.03g Tejido Sano: 0.14g	Lesión : 540,000 Tejido Sano: 560,000	Lesión:73% Sano:84%	Lesión: 270,000 Tejido Sano: 280,000
021	Adenocarcinoma intestinal invasor	Tumor: 0.03g Tejido Sano: 0.03g	Tumor: 240,000 Tejido Sano:207,000	Tumor:59% Sano:68%	Tumor: 104,000 Tejido Sano: 107,000

Cabe mencionar que las últimas dos muestras recibidas de cáncer gástrico del tipo difuso y del tipo intestinal además del tejido adyacente sano correspondientes a las muestras #34 y #35 respectivamente han sido congeladas con medio y DMSO sin disgregar ni fijar como las biopsias anteriores ya que se busca evaluar si llevar acabo estos procesos *a posteriori* permite una mejora en la viabilidad, una variable importante en el protocolo de Hi-C y fundamental en los experimentos de scRNA-seq a los que son sometidas estas muestras para obtener información integral a cerca de la topología de este cáncer y su expresión genética.

Experimento Piloto de Hi-C en células AGS

Para estandarizar el protocolo de Hi-C modificando los volúmenes necesarios para el bajo número celular frecuentemente obtenido de las biopsias y resecciones de cáncer gástrico (aproximadamente 1 millón o menos ver tabla 1). Se llevó a cabo el experimento de Hi-C en el cultivo de la línea celular AGS correspondiente a un adenocarcinoma gástrico del tipo difuso, siguiendo las condiciones descritas en materiales y métodos. Se contó 1 millón de células por hematocitómetro y estas fueron procesadas como se describe en el apartado del protocolo de Hi-C hasta el paso de extracción del ADN dónde se realizaron los controles que se describirán a continuación.

Controles de Hi-C

Control de Ligación

Después de la extracción del ADN en el experimento de Hi-C se realizan controles que permiten conocer si los pasos posteriores a la digestión (en este caso con la enzima de restricción Mbo1) se realizaron con éxito. Primero, se lleva acabo una electroforesis con el ADN obtenido del rellenado, marcaje con biotina y ligación. El peso molecular debe visualizarse de forma similar al tamaño del ADN sin digerir (Ver materiales y métodos). Como se muestra en la figura 15, el ADN obtenido de las células AGS sometidas a los primeros pasos del Hi-C presenta un alto peso

molecular (Figura 15, la banda en el carril 4 se señala con una flecha blanca), similar al de un control de genómico no digerido (Figura 15 gel A) donde además se observa un barrido característico de fragmentos quiméricos de menor tamaño. Este resultado muestra que estos pasos se han realizado correctamente. Los controles del ADN digerido y no digerido no pudieron ser observados en este experimento debido a su baja concentración, aunque fueron cargados en los carriles 2 y 3 de la figura 15.

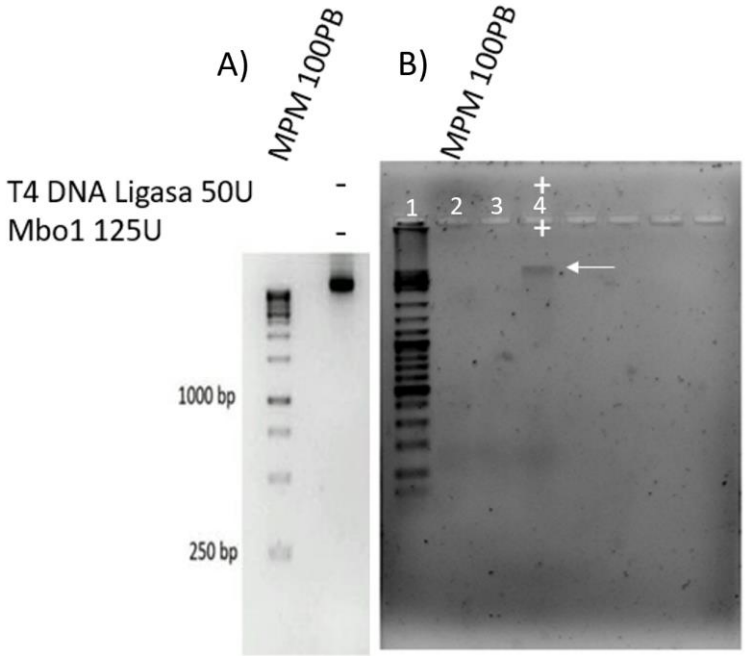


Figura 15. La etapa de ligación se realizó correctamente en el experimento con 1 millón de células AGS, control de ligación. A) Gel ideal de ADN sin digerir y ligar. **B)** Gel del material genético digerido con 125U de Mbo1 y ligado con 50U de T4 ADN Ligasa, señalado por una flecha blanca en el carril 4 de las células AGS.

Control Interno

El control interno consiste en realizar una PCR que amplifica el producto de ligación entre fragmentos contiguos de restricción (Ver anexo de tabla de primers, Figura 18 ”Control interno”). La ligación entre fragmentos contiguos o muy cercanos en el genoma lineal se lleva a cabo con alta frecuencia y representa un control de que la reacción de ligación ha sucedido de manera eficiente (Ver figuras 16 y 17).

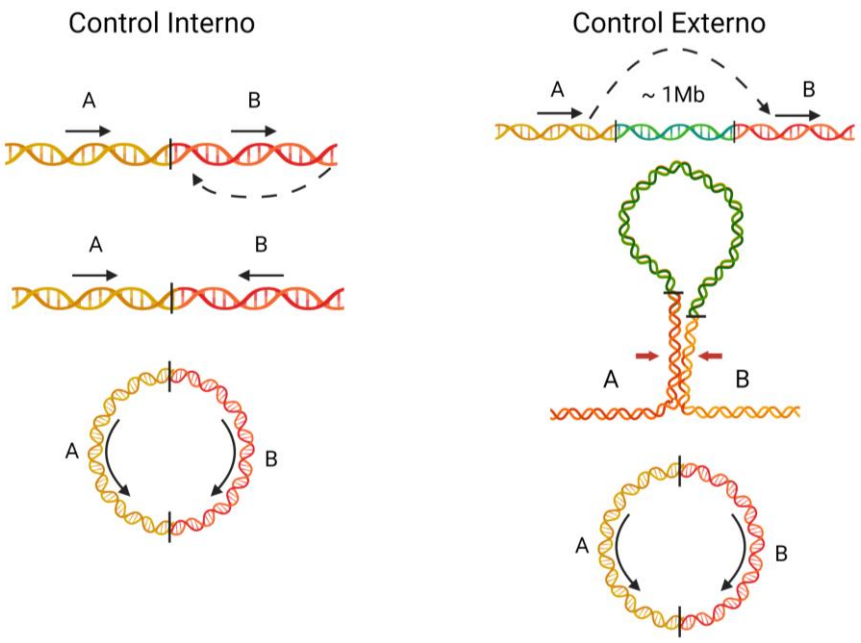


Figura 16. Diagrama de los controles interno y externo de Hi-C. Izquierda, el control interno busca recuperar una interacción conocida entre fragmentos de restricción contiguos que se explica por cercanía en secuencia. Derecha, el control externo está diseñado para recuperar una interacción lejana conocida que tiene relevancia biológica, mostrando que el experimento fue exitoso (Leyendas A y B representan *loci*, flechas negras indican la direccionalidad de las moléculas de ADN, flecha punteada indica interacción lejana y finalmente, las flechas rojas indican *loci* interactuando en una asa de cromatina. Figura generada en Biorender).

El amplicón del control interno diseñado en este experimento genera un producto de 186pb (Figura 18 carril 1 ”Int S/D” y diagrama A). Además, estos fueron digeridos con la enzima Mbo1 ya que el rellenado del sitio de corte con esta enzima genera 2 sitios de restricción para la misma (Figura 17 C rellenado y marcaje con biotina y D ligación) y a su vez se genera un sitio de restricción para la enzima Cla1 5´ATCGAT3´ 3´TAGCTA5´ (Figura 17 E diagrama del fragmento quimérico obtenido).

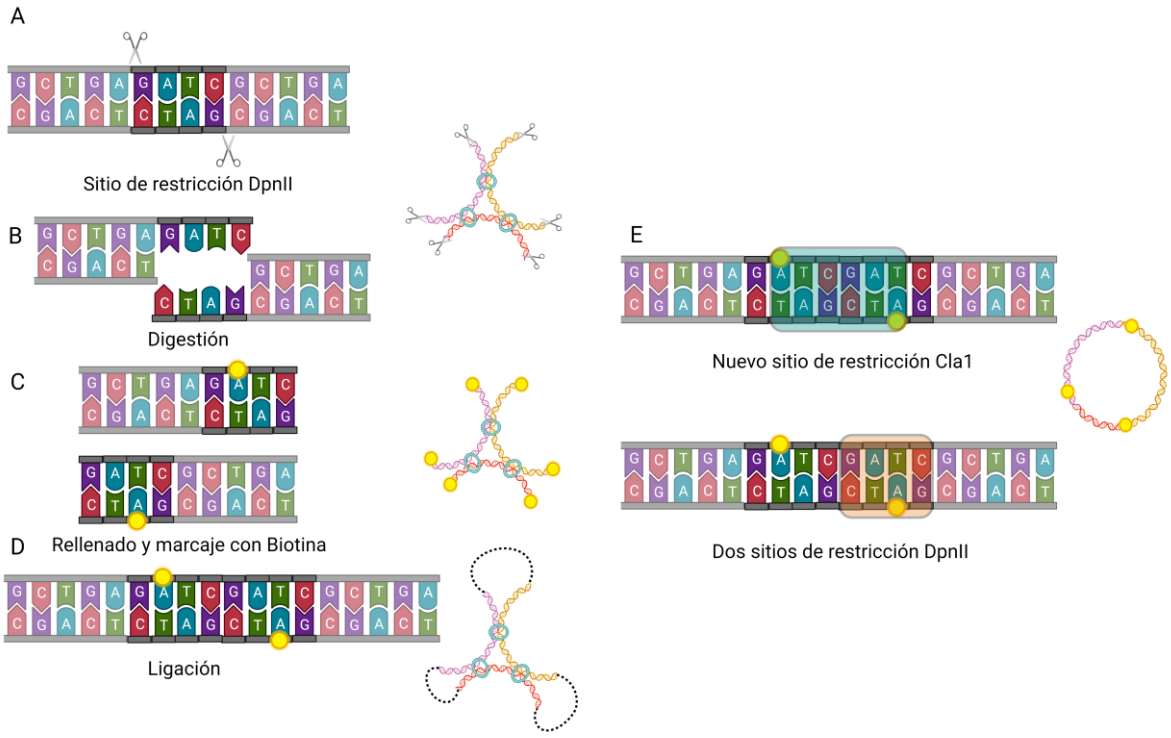


Figura 17. Diagrama de los nucleótidos involucrados en la digestión, rellenado y ligación en el protocolo de Hi-C usando la enzima de restricción DpnII. E) Se muestra la obtención del sitio de restricción de Cla1 al tener dos sitios de restricción de DpnII. La biotina de la adenina se muestra con una esfera amarilla. Cada inciso presenta el diagrama de cada uno de los pasos en el núcleo celular (figura generada en Biorender).

La digestión con Cla1 nos permite evaluar si el rellenado de nucleótidos y por lo tanto el marcaje con biotina se llevó a cabo de manera eficiente. Al digerir el amplicon con esta enzima se obtuvieron fragmentos de137 pb y 49 pb (Figura 18 gel de material de Hi-C de células AGS digerido Mbo1 ”Int Mbo1” y Cla1 ”Int Cla1” segundo y tercer carril respectivamente y diagrama A).

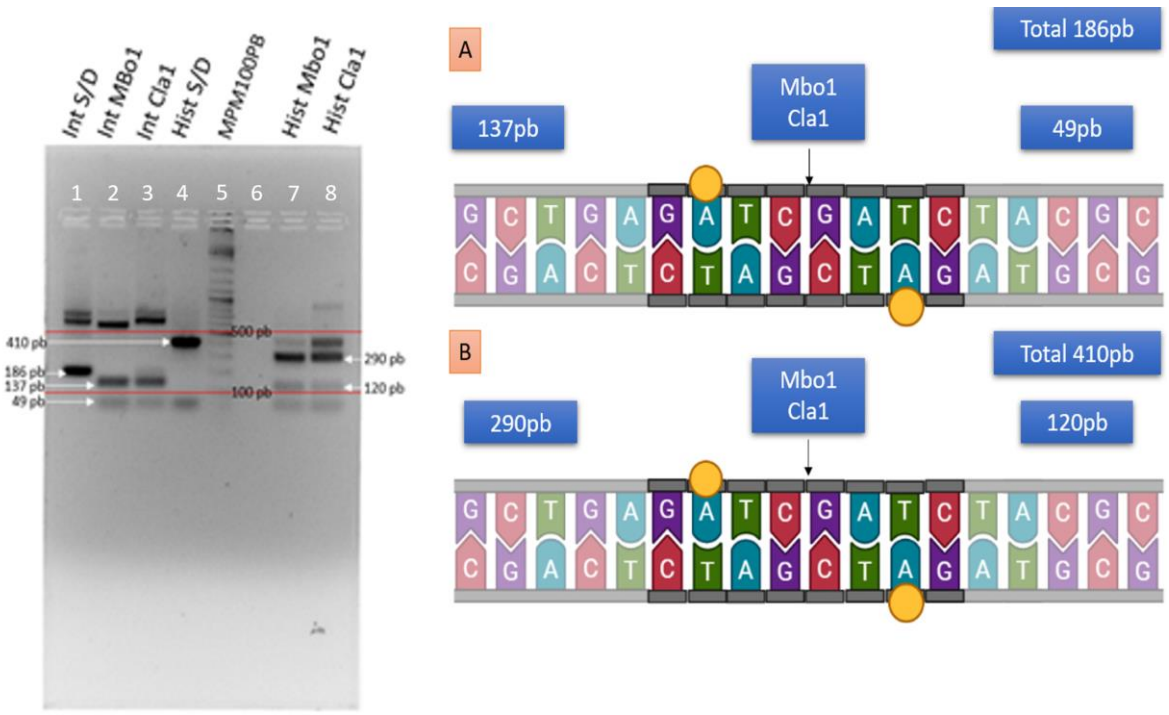


Figura 18. El experimento piloto de Hi-C en Células AGS ha sido digerido y ligado correctamente. Gel de electroforesis del material genético quimérico de las Células AGS amplificado para el control interno del lado izquierdo del gel carriles 1,2 y 3 dónde fue digerido con Mbo1 y Cla1 y del lado derecho la amplificación para un cluster de histonas del cromosoma 6 donde también se obtienen las digestiones esperadas por Mbo1 y Cla1 carriles 7 y 8, también se marcan las bandas relevantes con flechas blancas. Los diagramas A y B representan las longitudes de los fragmentos obtenidos por la digestión (Imagen generada en Biorender.com).

Por lo tanto, el rellenado con biotina y la ligación ocurrieron de manera correcta en este experimento.

Control de amplificación de una interacción de largo alcance

Para comprobar si con el material genético obtenido después de la digestión y ligación se pueden recuperar interacciones conocidas a larga distancia en el genoma, se llevó a cabo una PCR dónde se utilizaron oligos diseñados para amplificar productos de ligación entre los fragmentos de restricción que contienen a los genes de histonas (Ver anexo de tabla de primers, Figura 18 ”Control externo”), distribuidos en el cromosoma 6 a ~1Mb de distancia (Figura 16, Control Externo).

La amplificación de este producto de ligación genera un amplicón de 410 pb (cuarto carril en la figura 18 ”Hist S/D” y diagrama B). Al digerir el amplicón con Mbo1 y Cla1 se obtuvieron fragmentos esperados de 290 pb y 120 pb en los carriles séptimo y octavo de la figura 18 (Hist Mbo1 y Cla1 respectivamente) las bandas están marcadas con flechas blancas y el diagrama B representa los fragmentos obtenidos.

Por lo tanto, se concluye que el experimento fue exitoso, ya que se recuperan interacciones de largo alcance con relevancia biológica, no solo interacciones inherentes a la cercanía entre *loci* como se muestra en la sección anterior.

Una vez obtenidos los controles que confirman el rellenado de los fragmentos con biotina-dATP y la ligación de los fragmentos se continuó con el experimento.

Preparación de biblioteca para secuenciación

Brevemente, se llevaron a cabo los pasos de la sonicación de las moléculas para obtener fragmentos de DNA pequeños y que puedan ser secuenciados en una plataforma ilumina, el pulldown de biotina que nos permite tener un filtrado selectivo de solo los fragmentos quiméricos marcados con biotina-dATP y la preparación de las bibliotecas para secuenciación masiva (ver sección de materiales y métodos. Figura 10).

En este punto, como control se realiza un experimento de PCR con primers que reconocen las secuencias de los adaptadores Tru-seq (Ver Anexo “primers para adaptadores Tru-Seq”). El resultado esperado es un barrido característico que representa la colección de moléculas obtenidas. Sin embargo, aunque se repitió el experimento dos veces con la ligasa T4 de la casa comercial NEB y una vez se realizó con la ligasa T4 de la casa comercial ThermoScientific no fue posible amplificar la biblioteca. En el gel de la figura 19 se puede observar en los carriles 2 y 3 el barrido esperado, estas son librerías derivadas de otros experimentos exitosos utilizados como control, en los carriles 4 y 5 se colocó la biblioteca derivada del experimento realizado con 1 millón de células AGS donde no se observa el barrido esperado.

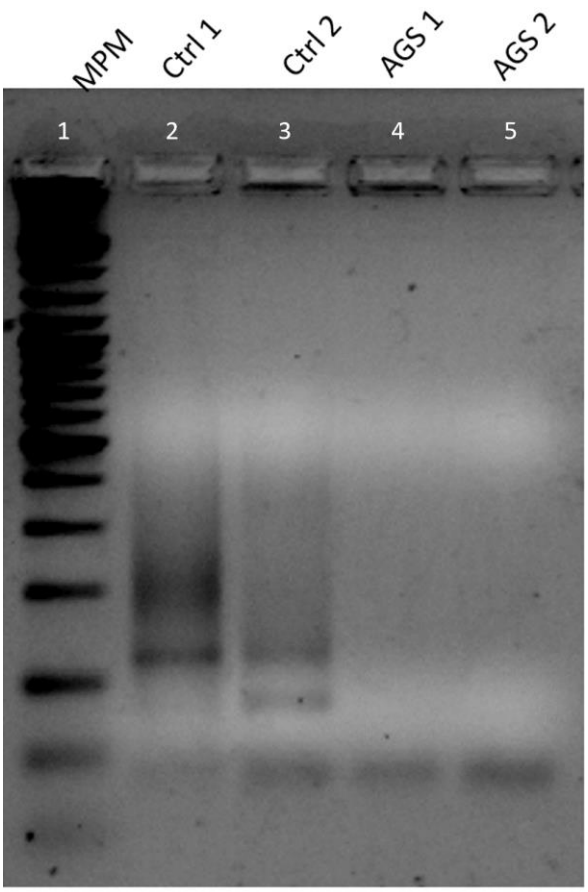


Figura 19. La amplificación con primers para adaptadores Tru-Seq del experimento piloto de Hi-C en Células AGS no se llevó acabo. Se observa el barrido esperado en los carriles 2 y 3 con librerías control, los carriles 4 y 5 correspondientes a la librería de AGS por duplicado no amplificaron, mostrando que la ligación de los adaptadores no fue exitosa.

En conjunto, este experimento piloto validó el éxito en las primeras etapas de Hi-C: fijación, digestión, rellenado y marcaje con biotina, ligación, purificación del ADN y sonicación, al lograr llevarse a cabo con una cantidad de ADN equivalente a la esperada en las muestras de pacientes. Esto respalda la correcta estandarización de estos pasos, incluyendo las concentraciones de reactivos utilizados en la técnica.

No obstante, durante las etapas finales de Pull-Down de biotina y preparación de la librería, se encontraron dificultades, como se ha explicado anteriormente (Ver figura 19). Por lo tanto se decidió llevar acabo un nuevo experimento.

Experimento de Hi-C con 5 millones de células AGS y K562

En el experimento con 5 millones de células AGS se utilizó como control un experimento igual en diseño que se realizó a la par con la línea celular de leucemia mieloide K562, a la cual se le ha realizado Hi-C en el laboratorio anteriormente de manera exitosa. Con la finalidad de descartar alguna particularidad de la línea celular AGS que afecte la eficiencia del Hi-C.

Como primer paso, se realizó una prueba de digestión de las enzimas Mbo1 y Dpnll sobre ADN genómico de la línea celular K562 para mostrar su efectividad, la cual fue la esperada para cada enzima al ver un barrido característico de la digestión del material genético por parte de estas enzimas (Ver Figura 20 gel A materiales digerido “D” carriles 1,3 y no digerido “ND” carriles 2,4 para cada enzima). Cabe destacar que el motivo de corte de estas enzimas es el mismo 5´…GATC…3´ 3´…CTAG…5.

Se llevó a cabo la digestión del material genético extraído de las células AGS y K562 con la enzima Dpnll que como se mencionó anteriormente tiene un sitio de corte igual a Mbo1. Finalmente, se corrió el gel correspondiente y se observó la eficiencia de la digestión esperada para ambos materiales tratados por Hi-C (Ver Figura 20 gel B, con los materiales digerido “D” y no digerido “ND” de las células AGS y K562). Por lo tanto se prosiguió con el protocolo de Hi-C utilizando la enzima DpnII.

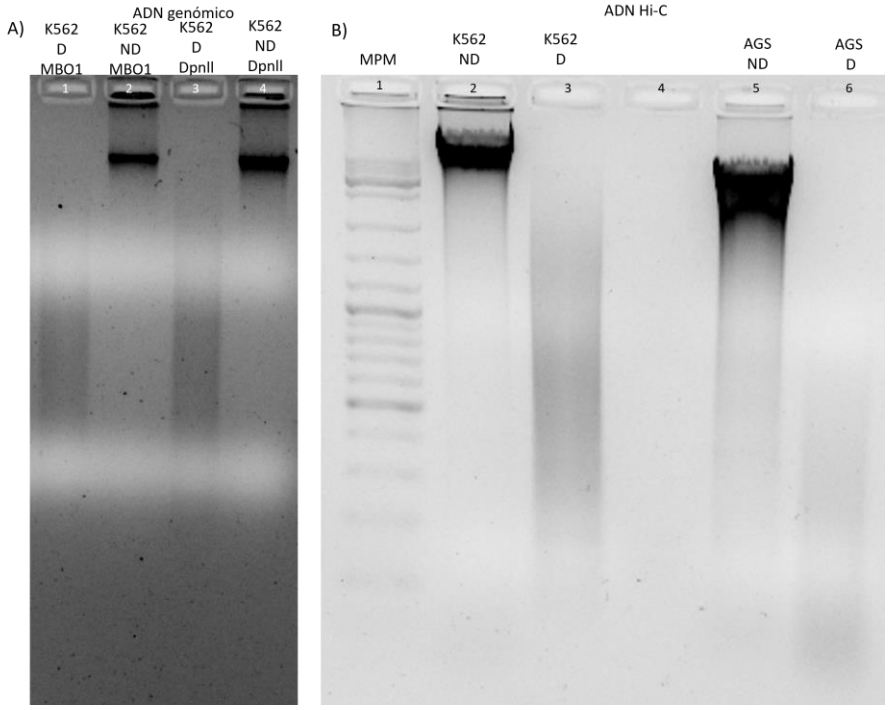


Figura 20. Las enzimas de restricción utilizadas para el protocolo de Hi-C son eficientes. A) Primer carril material digerido con Mbo1 y el no digerido en el carril 2, además del material digerido y no digerido del del ADN genómico de K562 con Dpnll, carriles 3 y 4 respectivamente. Se puede observar el barrido característico de la digestión en ambos casos comprobando que las enzimas funcionan correctamente. **B)** Material no digerido y digerido de la línea celular K562 carriles 2 y 3 y en carriles 5 y 6 material de células AGS ambas con el tratamiento de Hi-C, donde se puede observar la digestión eficiente esperada.

Posteriormente, se evaluaron los controles interno y externo como se ha mencionado anteriormente para cada uno de los experimentos además de digerir estas reacciones de PCR con la enzima Cla1 como se muestra en la figura 21 (Ver anexo de tabla de primers, Figura 21). Sin embargo para el caso de AGS no se obtuvieron los fragmentos esperados después de la digestión para el control interno de 290 y 229 pb pero se recuperó la banda de 519 pb mostrando una digestión ineficiente. Por otro lado, en el experimento control con K562 si se pudieron observar estas bandas de 290 pb y 229 pb para el control interno como se muestra en el carril 4 de la figura 16. En el caso de los controles de interacción externos se muestra el mismo fenómeno, no recuperando los fragmentos esperados de 274 y 290 pb en el caso del Hi-C de AGS (carril 3 de la figura 21) y recuperando los

fragmentos esperados en el caso de K562 de 278 y 229 pb, como se muestra en el carril 5.

Tomando estos resultados en conjunto se puede concluir que los pasos del digestión y ligación en el Hi-C de la línea celular AGS no ocurrieron correctamente por lo tanto no se siguió con el experimento.

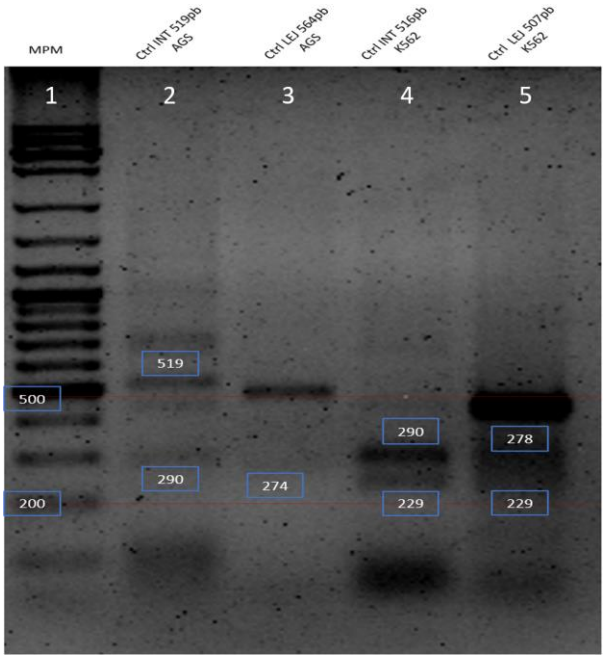


Figura 21. Experimento con cinco millones de células AGS y K562 controles interno y externo digeridos con Cla1. No se recuperan los fragmentos esperados en el Hi-C de las células AGS en ambos casos, carriles 2 y 3. Sin embargo en ambos controles para el Hi-C de K562 se observan los fragmentos esperados después de la digestión, carriles 4 y 5.

Debido a que este experimento y el anterior confirman que ya se tenían estandarizadas las etapas principales del Hi-C y tomados en conjunto muestran que la técnica era eficiente al ser realizada, se prosiguió con el experimento de Hi-C en una biopsia de cáncer gástrico difuso, el tejido adyacente al tumor y las PMBCs del mismo paciente.

Experimento de Hi-C en las muestras obtenidas del paciente 8. Adenocarcinoma gástrico del tipo difuso

Después de llevar a cabo los experimentos previos para estandarizar las etapas principales del Hi-C, se procedió a trabajar con una muestra de adenocarcinoma gástrico del tipo difuso, específicamente del paciente 8, cuyos detalles se encuentran en la tabla 1. Es importante destacar que la viabilidad del tejido adyacente y del tumor se estimó en un 80%, con un total de 1,900,000 y 3,200,000 células, respectivamente.

Debido a la cantidad de células en la muestra de tumor y tejido adyacente, se optó por utilizar las células mononucleares de sangre periférica (PMBCs) del mismo paciente para llevar a cabo los controles mencionados en la sección anterior durante la realización de este experimento. Las PMBCs presentaban una viabilidad calculada del 80% y un número celular de 2,000,000.

El objetivo principal de este trabajo es estandarizar la técnica de captura conformacional cromosomal de alto rendimiento Hi-C en muestras de pacientes con cáncer gástrico difuso. Las muestras del paciente 8 son ideales para este propósito, en primer lugar, debido al número y viabilidad celular adecuada del tejido adyacente, el tumor de cáncer gástrico difuso y las PMBCs, como se evidenció en los experimentos previos. Además, estas células provenientes de las mismas muestras se utilizaron también para el protocolo de scRNA-seq, donde la viabilidad puede ser un factor limitante (ver resultados en secciones posteriores). En conjunto, estos experimentos permitieron comprender la topología del genoma y su impacto en la expresión genética de las muestras mencionadas anteriormente.

Una vez realizados los pasos de fijación, digestión con DpnII, rellenado, marcaje con biotina y purificación del ADN (ver figura 10), se obtuvieron 500 ng de ADN para las PMBCs, 150 ng de ADN para el tejido sano y 200 ng de ADN para el tumor.

La prueba para observar el material genético ligado se llevó a cabo solo en el caso de PMBC como se muestra en el carril 8 de la figura 22 donde se observa la banda de alto peso molecular esperada con un barrido que representa la población de los fragmentos quiméricos obtenidos por la ligación. Los controles de digestión no se pudieron observar debido al poco material obtenido y cargado (carriles 3, 5, 6 para material No digerido “ND” y digerido “D DpnII” respectivamente).

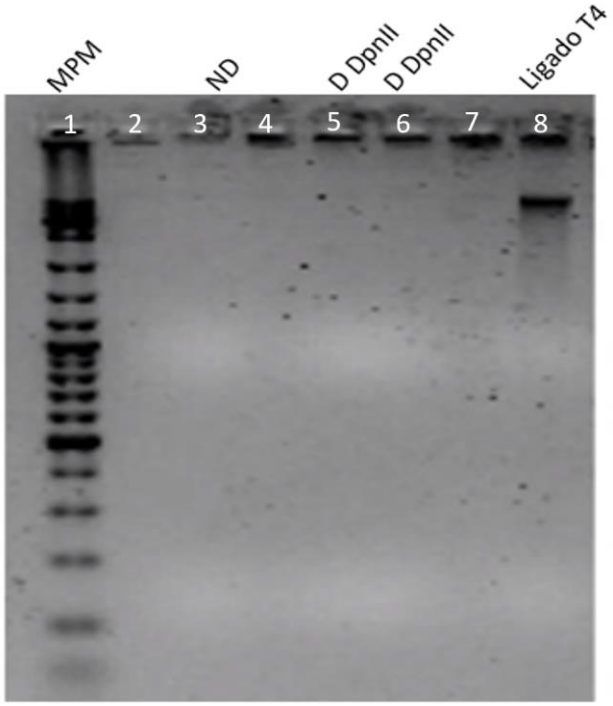


Figura 22. El experimento de Hi-C en las células PMBCs se ligó correctamente. Material no digerido “ND”, digerido“D DpnII” y ligado “Ligado T4” de las PMBC del paciente 8. No se observó el material digerido y no digerido debido a la concentración del material obtenido. Sin embargo, se recuperó el material de alto peso molecular con un barrido esperado después de la ligación en el carril 8.

Posteriormente se continuó con la sonicación y el pulldown de biotina para las tres muestras, además se les ligaron los adaptadores para secuenciación Tru-seq como parte de la preparación de la biblioteca, se comprobó que los primers para los adaptadores eran capaces de amplificar la biblioteca de las PMBCs (Ver Anexo

“primers para adaptadores Tru-Seq”) y se utilizó la misma para determinar el número de ciclos de amplificación que se utilizaron para las bibliotecas del tejido adyacente y el tumor (Figura 23).

El racional de esta etapa es que se utilizan el numero de ciclos donde la biblioteca se observe ligeramente en el gel de agarosa y se espera una colección de moléculas “barrido” entre 400 pb y 1000 pb, al elegir el menor número de siclos se evitan moléculas no informativas derivadas de la sobre amplificación de la biblioteca.

Amplificando la librería de PMBCs 6, 7, 8 y 9 ciclos, se determinó que los ciclos óptimos para todas las librerías (PMBCs, tejido adyacente y tumor) serían 7.

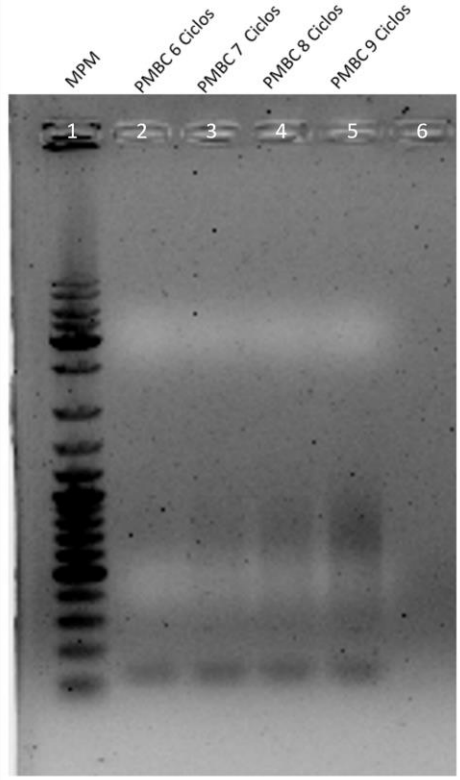


Figura 23. Determinación de ciclos de amplificación óptimos para la biblioteca de PMBCs. PCR de 6, 7, 8 y 9 ciclos de la librería de las PMBC. Esta amplificación permite determinar el número de ciclos óptimos para la amplificación de la librería evitando dúplicas de PCR quedándose con los ciclos donde se marca ligeramente la biblioteca, 7 ciclos en este caso. Este dato se utilizó para amplificar las librerías del tejido adyacente y tumor del paciente 8.

Una vez amplificadas todas las librerías se realizó una selección de tamaños de las moléculas mediante SPRI beads para obtener moléculas cercanas a 500pb que son del tamaño adecuado para la secuenciación.

Se tomaron 2µL de cada una de las librerías (PMBCs, tejido adyacente y tumor) se amplificaron y se digirió la mitad de la reacción de PCR con Cla1 esperando un cambió en el tamaño de las moléculas “barrido” hacia menores pesos moleculares, comprobando así que la población fue digerida y ligada correctamente (Ver sección de controles de Hi-C, Figura 17) como se muestra en la figura 24.

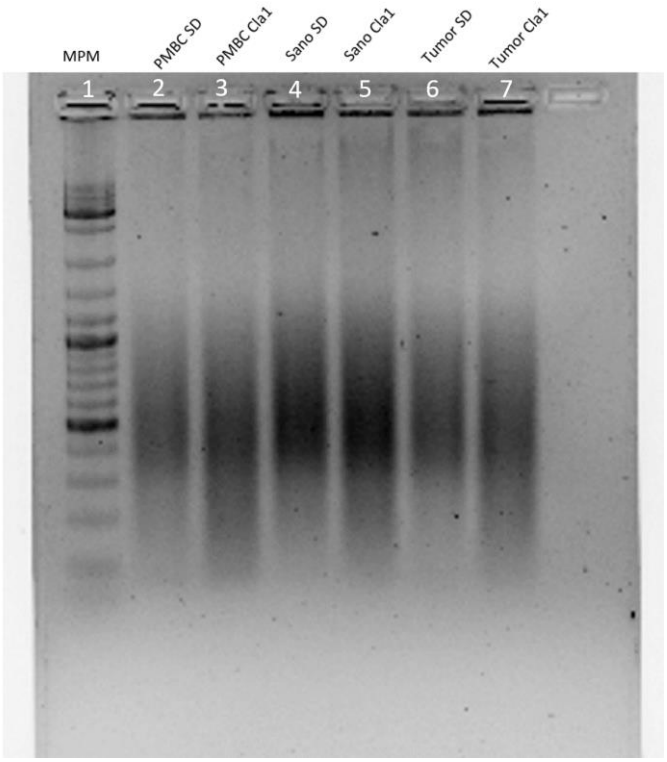


Figura 24. Control de digestión con Cla1 de las bibliotecas de Hi-C del paciente 8. En el gel se observan las librerías sin digerir (SD) y digeridas (Cla1) para las muestras correspondientes a PMBCs, Tejido Adyacente y Tumor del paciente 8 con adenocarcinoma difuso. En todos los casos se observa un corrimiento del barrido hacia pesos moleculares inferiores mostrando que la digestión procedió adecuadamente y en consecuencia reflejando el éxito del experimento (Marcador de peso molecular MPM carril 1 de 100 pb).

Finalmente, para confirmar el tamaño y concentración de las librerías una muestra de estas fue analizada por High Sensitivity D1000 ScreenTape obteniendo moléculas con una media de 484 pb para la librería de PMBC, 500 pb para el tejido sano y 417 pb para el tumor como se muestra en la figura 25. Estas concentraciones y tamaños son adecuados, por lo cual se procedió a secuenciar todas las muestras en una plataforma ilumina de forma pareada con lecturas de 150 pb.

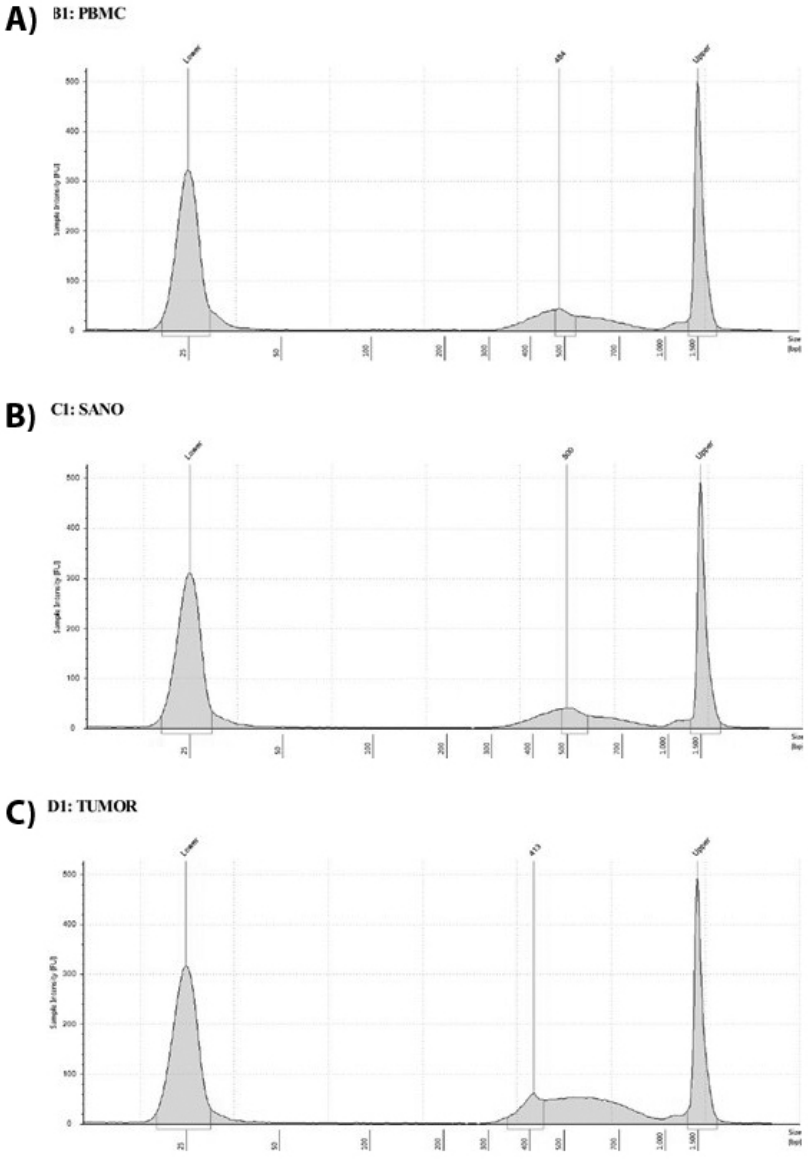


Figura 25. La selección de tamaños de las bibliotecas de Hi-C por SPRI beads fue optima. Tape de las librerías **A)** PMBC, **B)** tejido adyacente y **C)** tumor del paciente 8. Mostrando la concentración y el tamaño esperado para las mismas en el histograma central, mientras que los picos a ambos lados muestran los marcadores de peso molecular de 25 pb y 1500 pb, respectivamente.

Análisis Bioinformático

Control de calidad HiCUP

Las muestras del paciente 8 se secuenciaron obteniendo de manera pareada 548,022,870 de lecturas para PMBC, 547,416,878 para el tejido adyacente y 622,401,490 para el tumor.

Estos datos fueron analizados mediante el pipeline HiCUP[68] el cual permite distinguir los errores asociados a la secuenciación tales como las lecturas muy cortas para ser mapeadas, lecturas que se alinean múltiples veces o que no se alinean al genoma de referencia. En particular para las librerías de Hi-C existen múltiples lecturas que no son informativas, pero que son inherentes a la naturaleza del experimento y son identificadas mediante los sitios de restricción fusionados obtenidos después de la digestión, rellenado y ligación (ver figuras 10 y 17) las cuales serán descritas a continuación.

Primero, las lecturas informativas en los experimentos de Hi-C son denominadas en inglés como **uniqueDi-tags**[68] o **Interacciones de Hi-C**[74] dependiendo del autor y estas corresponden a los fragmentos obtenidos después de la fijación, digestión y ligación del genoma. Por lo tanto, deben contener la fusión de dos fragmentos de ADN diferentes, separadas por un sitio de restricción modificado como se muestra en la figura 17 en el índice E. Además, deben estar separados por una distancia genómica lineal relevante en *cis* o *trans* para poder atribuir su cercanía espacial a la organización tridimensional del genoma y no solo a su proximidad en secuencia (diagrama en la Figura 26 A).

De manera contraria los **fragmentos de religación** son consecuencia de la cercanía, representando interacciones entre regiones genómicas contiguas detectadas por HiCUP mediante los sitios de restricción modificados, siendo productos de religación que no aportan información estructural (diagrama en la Figura 26 B).

Otros artefactos encontrados en las librerías de Hi-C incluyen a los **fragmentos circulares** se generan a partir de una región digerida que se ligó sobre si misma pero que es recuperada en la librería mediante el pull-down ya que fue reparado con adenina biotinilada durante la etapa de rellenado (diagrama en la Figura 26 C). Sin embargo, HiCUP los descarta ya que solo mapean a un *locus*.

Este artefacto tiene como causa un error en la fijación del genoma, aumentando la distancia entre sitios de restricción lo que disminuye la probabilidad de las moléculas de ADN digeridas de ligarse con otras más lejanas[68].

Además, los fragmentos digeridos pero no ligados con otro *locus* pueden unirse con los adaptadores de secuenciación representando otra población no informativa en las librerías de Hi-C estos pueden ser identificados por HiCUP ya que algunos contienen un sitio de restricción y se denominan **Extremos Colgantes** (*Dangling end* en inglés) (Figura 26 D) y están relacionados con una remoción ineficiente de biotina. Por el contrario los fragmentos que no contienen sitios de restricción se les nombra **Fragmentos Internos** los cuales teóricamente no contienen la adenina biotinilada ya que esta se incorpora en los sitios de restricción (Figura 26 E, ver figura 10) una alta proporción de estos sugiere un pull-down ineficiente o la digestión en sitios no canónicos (Figura 10, ver Materiales y métodos, Protocolo de HiC).

Algunas interacciones pueden estar sobre representadas debido a **dúplicas de PCR** (Figura 26 F) HiCup las detecta ya representan moléculas que comprenden los mismos *loci* y con la misma extensión, esto está relacionado con los ciclos de PCR utilizados en la amplificación de la librería (Ver figura 23).

Por último, la validación del tamaño de las interacciones de Hi-C se lleva a cabo colocando los di-tags alineados en el genoma de referencia digerido *in-silico,* lo cual permite calcular la longitud teórica de las interacciones de Hi-C y elimina aquellas que salen del rango esperado, llamadas aquí **wrong size** (Figura 26 G).

Las explicaciones para estas discrepancias de tamaño incluyen di-tags con múltiples fragmentos internos, extremos colgantes o la pérdida o ganancia de sitios de restricción en la muestra biológica que no se encuentran en el genoma de referencia[68].

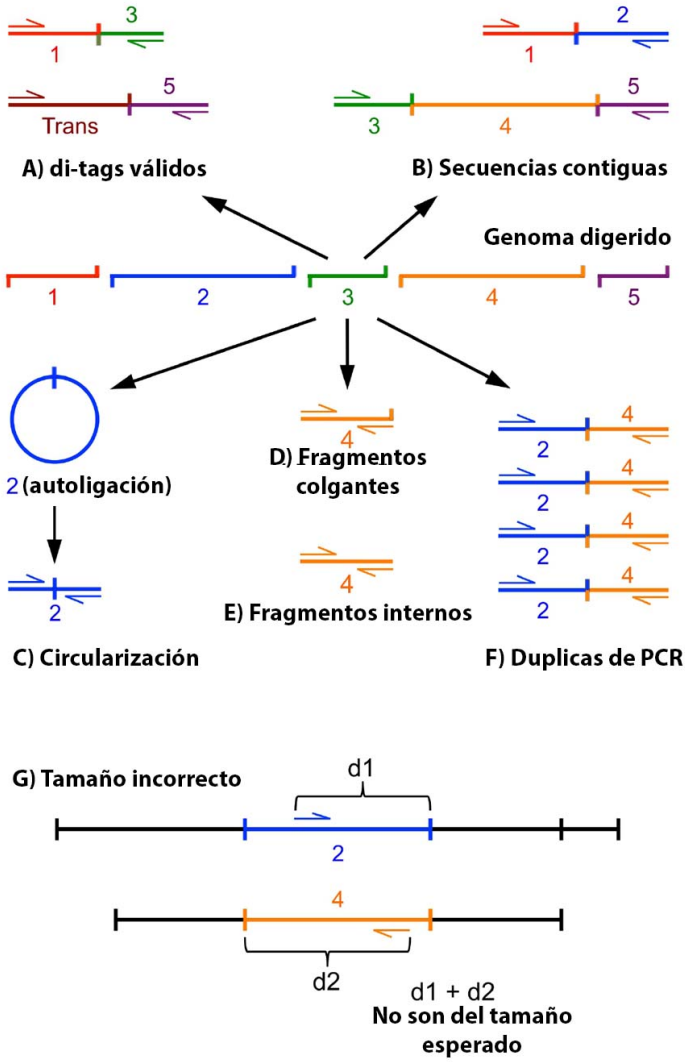


Figura 26. Diagramas de la diversidad de moléculas encontradas por HiCup en las librerías de Hi-C partiendo de un fragmento de genoma digerido. A) di-tags válidos o **unique di-tags** moléculas informativas de Hi-C. **B) fragmentos de religación** entre secuencias contiguas. **C) fragmentos circulares** ligados sobre si mismos. **D) Extremos colgantes** o dangling ends, fragmentos no ligados con otros locus que contienen un sitio de restricción y fueron secuenciados. **E) fragmentos internos** que no contienen sitio de restricción pero no son eliminados por el pull-down. **F) dúplicas de PCR** derivadas de la amplificación de la librería. **G)** lecturas de tamaño equivocado o **wrong size** que se desvían del tamaño teórico esperado por ganancia o pérdida de sitios de restricción en la muestra biológica (Modificada de Wingett *et al.* 2015[68]).

Comparación entre librerías de Hi-C de CG publicadas

Además de los datos de Hi-C de los tejidos del paciente 8, también fueron analizados los datos de Hi-C generados por el grupo de Ooi et al. en 2020 que comprenden la línea celular de carcinoma gástrico humano SNU16 y un tumor de cáncer gástrico del tipo intestinal, al que el grupo denominó T2000877.

El racional de este análisis es la comparación de experimentos equivalentes (Bibliotecas de Hi-C de PMBCs, tejido adyacente y tumor de CG difuso del paciente 8 con las bibliotecas de Hi-C SNU-16 y T2000877) en muestras biológicas de esta patología. Asimismo, siendo el trabajo Ooi el único reportado con estas características se vuelve el referente en distintas dimensiones, por mencionar alguna, el desempeño de la obtención de información topológica mediante el protocolo de Hi-C en tumores de cáncer gástrico, estas comparaciones se mostrarán en lo sucesivo.

Reportes de HiCUP

A continuación, se mostrarán y compararán los resultados obtenidos de los reportes de HiCUP realizados sobre los datos de Ooi SNU16 y T200087 y para los tejidos del paciente 8 PMBCs, tejido adyacente y Tumor.

Biblioteca de Hi-C, SNU16

Como se menciona anteriormente SNU16 es una línea celular de carcinoma gástrico humano. En este trabajo se tomó como punto de comparación la biblioteca de Hi-C SNU16 generada por Ooi *et al.* 2020[7] ya que presenta datos con poca presencia de artefactos de Hi-C, hablando de una alta eficiencia en el experimento. Cabe mencionar que de todos los experimentos de Hi-C aquí analizados es el único que no pertenece a un tejido, por lo tanto se espera que la viabilidad y el número celular no hayan sido factores limitantes para llevarse a cabo el protocolo de Hi-C.

En este experimento previamente publicado se secuenciaron 322,952,414 de lecturas pareadas, de las cuales las lecturas no truncadas, que corresponden a alineamientos únicos y pareados las cuales son las lecturas utilizadas para los posteriores análisis son 228,436,024 de estas 210,838,375 de lecturas son pares válidos de Hi-C (Figura 27) con solo un 2.5% de religación y un 2.1% de fragmentos internos como artefactos de Hi-C con mayor representación en la librería lo cual muestra cierta presencia de estas moléculas incluso en un experimento eficiente (figura 23).

Por último, 170,347,914 de lecturas pertenecen a Unique-Ditags un 80% de los pares válidos, los cuales están compuestos por 13.8% de interacciones en *cis* cercanas <10 kbp, 74.0% de interacciones en *cis* lejanas >10 kbp y finalmente 12.2% de interacciones en *trans*. Ejemplificando el comportamiento general de las librerías de Hi-C fijadas con formaldehido donde se espera un mayor número de interacciones en *cis* lejanas, uno menor de *cis* cercanas y por último las menores interacciones en trans (Figura 29).

Biblioteca de Hi-C, T200087

Una comparación más cercana en origen y condiciones a los tejidos del paciente 8 es la biblioteca de Hi-C del tumor de cáncer gástrico T200087 generada por Ooi *et al.* 2020[7] y previamente publicada, cuya librería de Hi-C al ser analizada con HiCUP contuvo 337,626,389 de lecturas pareadas, de las cuales 107,402,087 corresponden a lecturas no truncadas, alineamientos únicos y pareadas (el subgrupo necesario para continuar con el análisis) siendo el 55.4% de las lecturas totales. Se detectaron 107,402,087 de pares válidos para Hi-C (Figura 27).

Los artefactos de Hi-C para los datos del tumor T200087 ocuparon una parte importante de los pares válidos. Siendo los más relevantes en magnitud la religación con 16.5%, los fragmentos internos con 10.5%, los fragmentos de tamaño equivocado o wrong size con 6.9% y por último un 6.0% perteneciente a la religación

entre fragmentos contiguos (Figura 28). En conjunto, estos datos hacen referencia a la etapa de fijación y pull-down con el porcentaje de religación, fragmentos contiguos y fragmentos internos respectivamente. Además, las lecturas de tamaño equivocado pueden relacionarse con nuevos sitios de digestión no considerados en el genoma de referencia derivados de la inestabilidad genómica de este tumor[4,9].

Sin embargo, más allá de los errores que se pudieron haber suscitado durante el desarrollo de este protocolo y tomando en cuenta los datos que se mostrarán posteriormente en la sección dedicada al análisis de calidad de la librería de Hi-C del tejido sano del estómago del paciente 8 de la cual se conoce el número y la viabilidad de la muestra se hace la observación de que los artefactos encontrados en estas dos librerías T200087 y tejido sano pueden estar relacionados con el número celular y viabilidad de manera más importante que con errores en el protocolo, sobre todo considerando que las librerías de Hi-C generadas para los tejidos del paciente 8 fueron realizadas con idéntico protocolo en pasos y concentraciones en particular para este punto en la fijación con formaldehido y el pull-down de las moléculas biotiniladas.

Finalmente, el 82.15% de los pares validos 88,230,471 de lecturas corresponden a uniqueDi-tags los cuales incluyen un 22.7% de interacciones en *cis* cercanas <10 kbp, 59.7% de interacciones en *cis* lejanas >10 kbp y 17.7% de interacciones en *trans* (Figura 29).

Biblioteca de Hi-C, PMBCs

En las células PMBC del paciente 8 se obtuvieron 548,022,870 de lecturas pareadas, de las cuales 384,238,880 corresponden a lecturas no truncadas, alineamientos únicos y pareados las lecturas útiles para en análisis de los experimentos de Hi-C. Representando el 70% de éstas (Figura 27). 289,050,601 de lecturas pertenecen a pares válidos de Hi-C una vez filtrados de los artefactos anteriormente mencionados.

Los artefactos más representados en la librería corresponden a la religación de fragmentos cercanos con 13.6% y religación de fragmentos contiguos con 7.4% de las lecturas pareadas respectivamente (Figura 28), esto está relacionado con el tiempo de fijación del material genético como de describe en la sección anterior. Sin embargo, la eficiencia es suficiente para la obtención de datos topológicos.

Una vez eliminadas las dúplicas de PCR se obtuvieron 211,814,183 de uniqueDi-tags siendo el 73.28% de las lectura válidas para Hi-C conteniendo un 24.5% de interacciones en *cis* cercanas <10 kbp, 42.1% de interacciones en *cis* lejanas >10 kbp y finalmente 33.4% de interacciones en *trans* (Figura 29).

Biblioteca de Hi-C, Tejido adyacente sano

En el tejido sano del paciente 8 se secuenciaron 547,416,878 de lecturas pareadas, de las cuales el 61.1% son lecturas pareadas y 219,369,249 de lecturas son pares validos de Hi-C (Figura 27).

Respecto a los artefactos de Hi-C se encontró un 11.7% de religación, 8.6% de fragmentos internos, 6.1% de fragmentos circularizados y un 3.5% de extremos colgantes (Figura 28). Si bien estos artefactos de Hi-C están relacionados a cierta ineficiencia en la ligación y durante el pull-down de biotina[68]. Cabe mencionar que el protocolo seguido para el procesamiento de este tejido fue el mismo para las PMBC y el tumor en las cuales no se observa una prevalencia importante de fragmentos internos, que como se menciona anteriormente están relacionados con ineficiencia en el pull-down al mantener en la librería moléculas de ADN sin adenina biotinilada o bien por digestión en sitios no canónicos donde la adenina biotinilada sí es incorporada.

Realizando la comparación con el experimento del grupo de Ooi, el tumor de cáncer gástrico T2000877[7] se observa en este la presencia de fragmentos internos que comprenden un 10.5% de los pares válidos. Por lo tanto una posible explicación es

que estas secuencias estén con la inestabilidad genómica de los tumores de cáncer gástrico. Sin embargo, hacen falta más experimentos de este tipo para apoyar esta observación.

Finalmente, los Unique Di-Tags para el tejido sano del paciente 8 son 146,983,405 de lecturas representando el 67% de los pares válidos. Estos están conformados por un 28.8% de interacciones en *cis* cercanas <10 kbp, 43.8% de interacciones en *cis* lejanas >10 kbp y 27.4% de interacciones en trans (Figura 29).

Biblioteca de Hi-C, Tumor

En el tumor de cáncer gástrico del tipo difuso del paciente 8 fueron secuenciadas 622,401,490 de lecturas pareadas, 66.6% de estas son pareadas y 331,454,276 de lecturas resultaron pares válidos de Hi-C (Figura 27).

Respecto a los artefactos los principales en este experimento fueron la religación con 8.4% y los fragmentos circulares con 5.5% que refieren a la etapa de fijación (Figura 28).

Pese a esto, el número de Unique-Ditags corresponde a 238,862,489 de lecturas de las cuales el 21.7% son interacciones *cis* cercanas <10 kbp, 46.6% de interacciones en *cis* lejanas >10 kbp y 31.7% de interacciones en *trans* (Figura 29).

En conjunto, son datos suficientes para el análisis topológico. Cabe resaltar que en general las bibliotecas del paciente 8 presentan más interacciones en trans si se les compara con las bibliotecas de Ooi. Aunque, se mantiene la generalidad de una mayor cantidad de interacciones en cis que en trans en ambos sets de datos, este resultado puede estar relacionado con la viabilidad de las células del paciente 8 ya que la muerte celular puede enriquecer las interacciones en *trans*[18].

	SNU16	T2000877	PMBC	Tejido Sano	Tumor
Lecturas totales	322,952,414	337,626,389	548,022,870	547,416,878	622,401,490
Lecturas pareadas	228,436,024	186,964,793	384,238,880	334,215,082	414,308,309
Pares válidos	210,838,375	107,402,087	289,050,601	219,369,249	331,454,276
Unique-Ditags	170,347,914	88,230,471	211,814,183	146,983,405	238,862,489
Porcentaje de Unique Di-Tags	80.80%	82.15%	73.28%	67.00%	72.6%

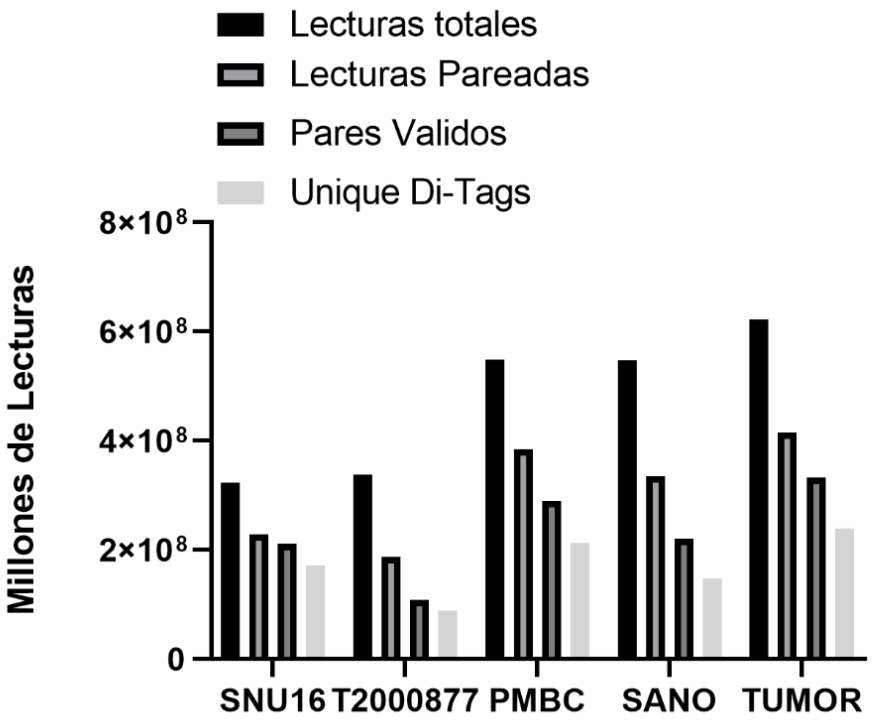
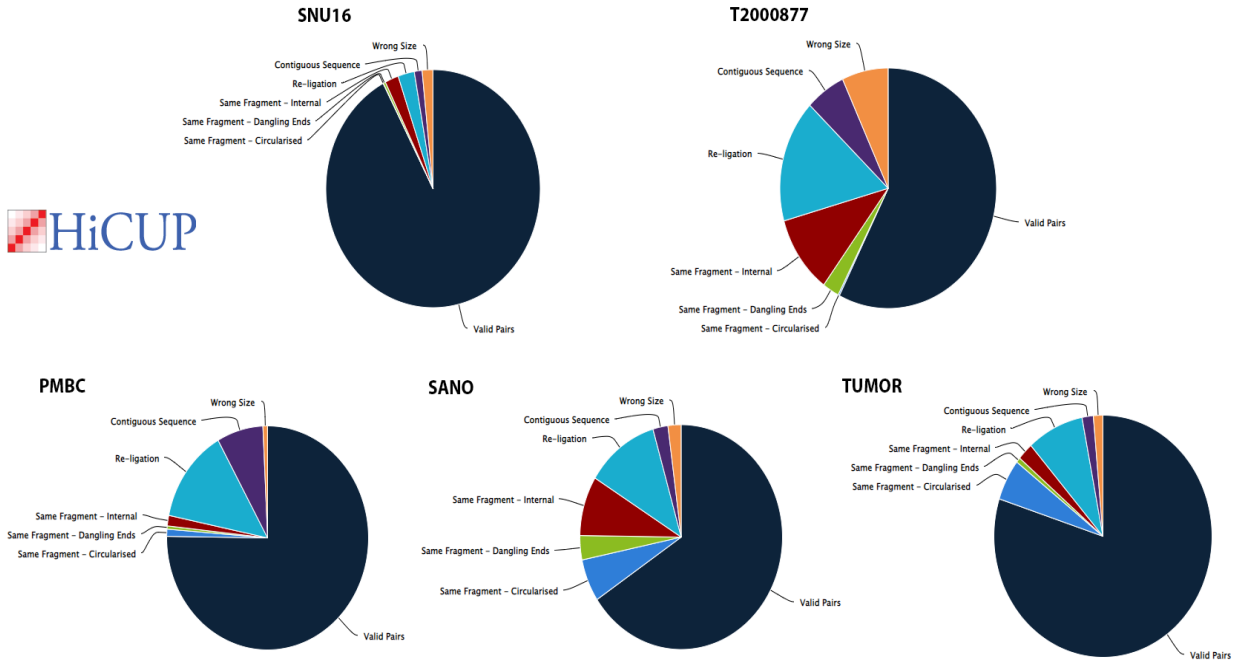


Figura 27. Reporte de lecturas para las librerías de Hi-C generado por HiCUP, datos del grupo de Ooi y tejidos del paciente 8. Arriba, tabla de **Lecturas totales** secuenciadas para cada librería. **Lecturas pareadas** donde se engloban las lecturas de alineamientos únicos, no truncadas y pareadas. **Pares válidos**, es decir lecturas de Hi-C detectadas por un sitio de restricción modificado uniendo dos fragmentos de ADN que mapean a dos locus distintos, además filtradas de los artefactos inherentes al protocolo. **Unique Di-tags**, lecturas de Hi-C filtradas de las dúplicas de PCR, mostrando el porcentaje de estas respecto a los pares válidos en la última fila de la tabla. **Abajo,** comparación de cada rublo entre las librerías analizadas mostrando la disminución de lecturas después del filtrado hasta los Unique Di-tags que serán las lecturas útiles para los análisis posteriores.



	SNU16	T2000877	PMBC	Tejido Sano	Tumor
Pares válidos	210,838,375	107,402,087	289,050,601	219,369,249	331,454,276
Religación	5,653,217 2.5%	30,841,604 16.5%	52,344,552 13.5%	39,056,526 11.7%	34,857,587 8.5%
Religación de fragmento contiguo	2,681,300 2.1%	11,130,269 6.0%	28,491,125 7.4%	8,062,586 2.4%	6,947,532 1.7%
Fragmentos circulares	106,853 0.0%	409,775 0.2%	3,971,161 1.0%	20,408,220 6.1%	22,825,366 5.5%
Fragmentos Internos	4,694,866 2.1%	19,582,700 10.5%	5,805,734 1.5%	28,708,406 8.6%	9,637,122 2.3%

Figura 28. Porcentaje de artefactos de Hi-C en cada librería analizada respecto a las lecturas pareadas. Arriba, gráficos de pastel de la representación de cada artefacto. Se observa una clara presencia de religación y religación de fragmento contiguo (en azul claro y morado respectivamente) en todos los tejidos a diferencia de la línea celular SNU16. Ambos artefactos están relacionados con las etapas de fijación y ligación del material genético. Sin embargo, su común presencia en mayor medida en tejidos, se puede relacionar con una característica propia de estos. Resaltando un porcentaje menor en las muestras del paciente 008 en general. **Abajo**, tabla de los artefactos más relevantes para los datos analizados y su porcentaje de representación. Cabe destacar, que la mayoría de los artefactos presentes lo están también en el tumor de cáncer gástrico T2000877 poniendo las librerías de los tejidos del P008 dentro de los resultado esperados. Excepto por la circularización que puede estar relacionada con la viabilidad y el número celular de la muestras del tejido sano y tumor (ver fila 4).

All Di-Tags	SNU16	T2000877	PMBC	Tejido Sano	Tumor
All Di-Tags	210,838,375	107,402,087	289,050,601	219,369,249	331,454,276
cis cercanas <10	29,193,386	24,333,751	70,855,501	63,099,032	71,792,848
kbp	13.8%	22.7%	24.5%	28.8%	21.7%
cis lejanas	156,042,259	64,148,579	121,813,040	96,163,594	154,650,460
>10 kbp	74.0%	59.7%	42.1%	43.8%	46.7%
Trans	25,602,730	18,919,757	96,382,060	60,106,623	105,010,968
	12.1%	17.6%	33.3%	27.4%	31.7%

Unique Di-Tags	SNU16	T2000877	PMBC	Tejido Sano	Tumor
Unique Di-Tags	170,347,914	88,230,471	211,814,183	146,983,405	238,862,489
cis cercanas <10	23,583,690	19,984,214	51,863,094	42,297,661	51,769,850
kbp	13.8%	22.7%	24.5%	28.8%	21.7%
cis lejanas	125,990,922	52,665,148	89,193,625	64,345,366	111,341,825
>10 kbp	74.0%	59.7%	42.1%	43.8%	46.6%
Trans	20,773,302	15,581,109	70,757,464	40,340,378	75,750,814
	12.2%	17.7%	33.4%	27.4%	31.7%

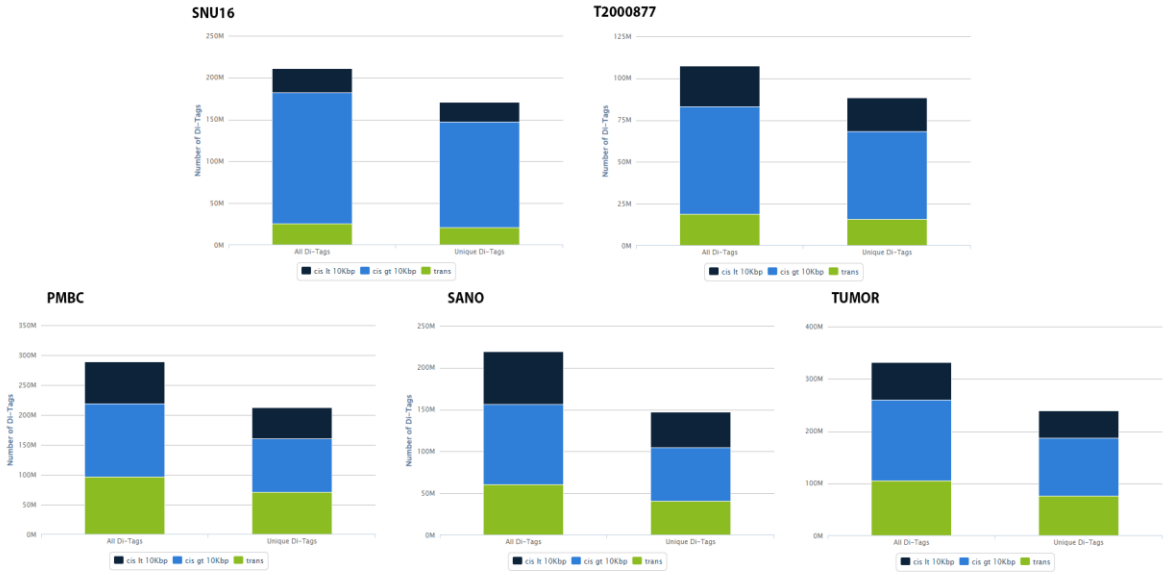


Figura 29. Composición de las interacciones de Hi-C encontradas en cada librería analizada respecto a la distancia genómica. Arriba, tablas de interacciones considerando las All Di-Tags, es decir todas las interacciones de Hi-C sin restar las dúplicas de PCR generadas en la amplificación de la librería (ver figura 18), estas están agrupadas en tres categorías: Interacciones en cis cercanas <10 kbp. Interacciones en cis lejanas >10 kbp e interacciones entre cromosomas, trans. Además, los Unique Di-Tags (lecturas que no consideran las dúplicas de PCR) separadas en las categorías mencionadas anteriormente. Ambas tablas contienen los porcentajes de representación de cada categoría en contraste de las All Di-Tags y los Unique Di-Tags respectivamente. **Abajo,** gráficos de barras de la presencia de cada categoría interacciones en cis cercanas <10 kbp (azul oscuro), cis lejanas >10 kbp (azul claro) y en trans (verde). El grueso de las interacciones son cis lejanas >10 kbp en todas las librerías lo cual es el comportamiento esperado. Sin embargo, en los tejidos del paciente 8 se observa una mayor cantidad en interacciones en cis cercanas <10 kbp y en trans respecto a los resultados de SNU16 y T2000877. Esto puede estar relacionado con la integridad de los mismos y el número celular.

Resultados de HiNT

Paciente 8 CG difuso

Se obtuvieron con HiNT-Pre las matrices de Hi-C de las librerías PMBCs, tejido adyacente y tumor correspondientes al paciente 8. Además, los datos de Ooi pertenecientes a la línea celular SNU16 y el tumor T2000877. Todas ellas visualizadas en Juicebox[72]. En general, no se encontraron cambios evidentes en las matrices de los tejidos del paciente 8. Es decir, las matrices correspondiente a PMBCs, tejido adyacente sano y tumor son equivalentes en tanto a VE (Figura 30). Cabe destacar que HiNT considera las lecturas normales de Hi-C, las quiméricas no ambiguas y las ambiguas, por lo tanto es capaz de detectar las translocaciones con enriquecimientos de lecturas entre cromosomas (*trans*) que son observadas en las matrices (Ver sección anterior). Sin embargo tanto la muestra de tumor de cáncer gástrico del tipo difuso como su control de tejido adyacente y PMBCs no muestran la presencia de translocaciones (Figura 30).

Al analizar estos datos mediante HiNT-CNV no se detectó variación en el número de copias llamándolas a 50kb (resolución estándar para este análisis), manteniendo la cobertura dentro del rango esperado de Log2CopyRatio (-0.3,0.3) por lo tanto, no existe pérdida o ganancia de material genético en esta muestra tumoral en particular y sus controles. En los diagramas correspondientes a cada tejido (Figura 31 A) PMBCs, B) Tejido adyacente sano y C) Tumor) se observan pequeñas regiones (puntos rojos) en algunos cromosomas que podrían ser interpretados como ganancias de material genético. Sin embargo, estas coordenadas fueron revisadas en las matrices y todas corresponden a regiones centroméricas.

Relacionado con la falta de variación en el número de copias, el módulo HiNT-TL tampoco detectó translocaciones, siendo coherente con lo observado en las matrices. Esta estabilidad genómica es característica de algunos tumores de cáncer gástrico del tipo difuso[4,9,67].

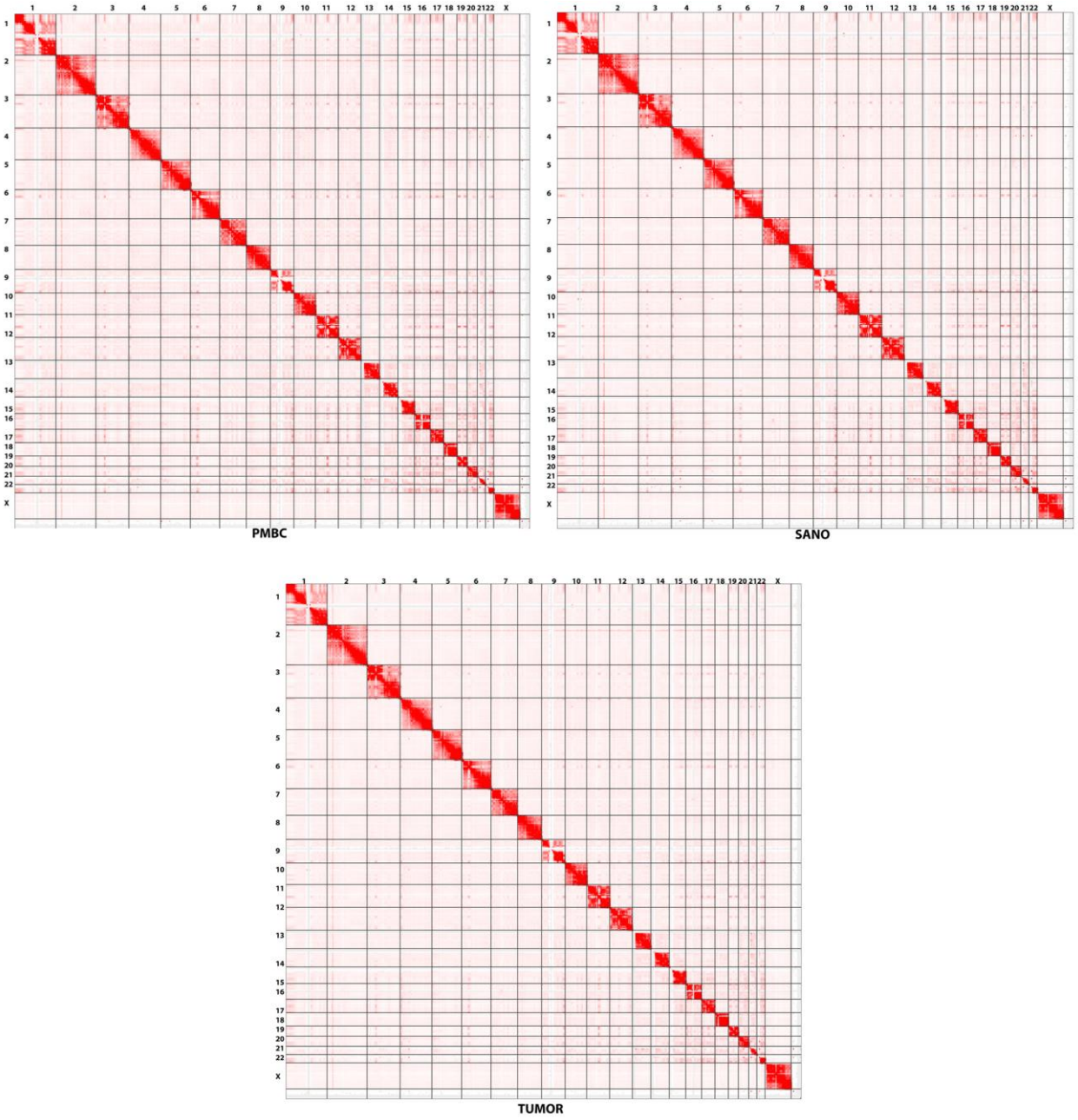


Figura 30. Matrices de Hi-C generadas por HiNT-pre correspondientes a los tejidos del paciente 8 PMBCs, tejido sano y tumor. En estas no se presentan interacciones ectópicas entre cromosomas (*trans*) en ninguno de los tejidos, a esta escala macro las matrices parecen equivalentes, cabe mencionar que este módulo de HiNT considera las lecturas normales de Hi-C, las quiméricas no ambiguas y las ambiguas al generar la matriz (En cada matriz los cromosomas están alineados del 1 al 22 y finalmente el cromosoma X correspondiendo cada uno a las filas y las columnas).

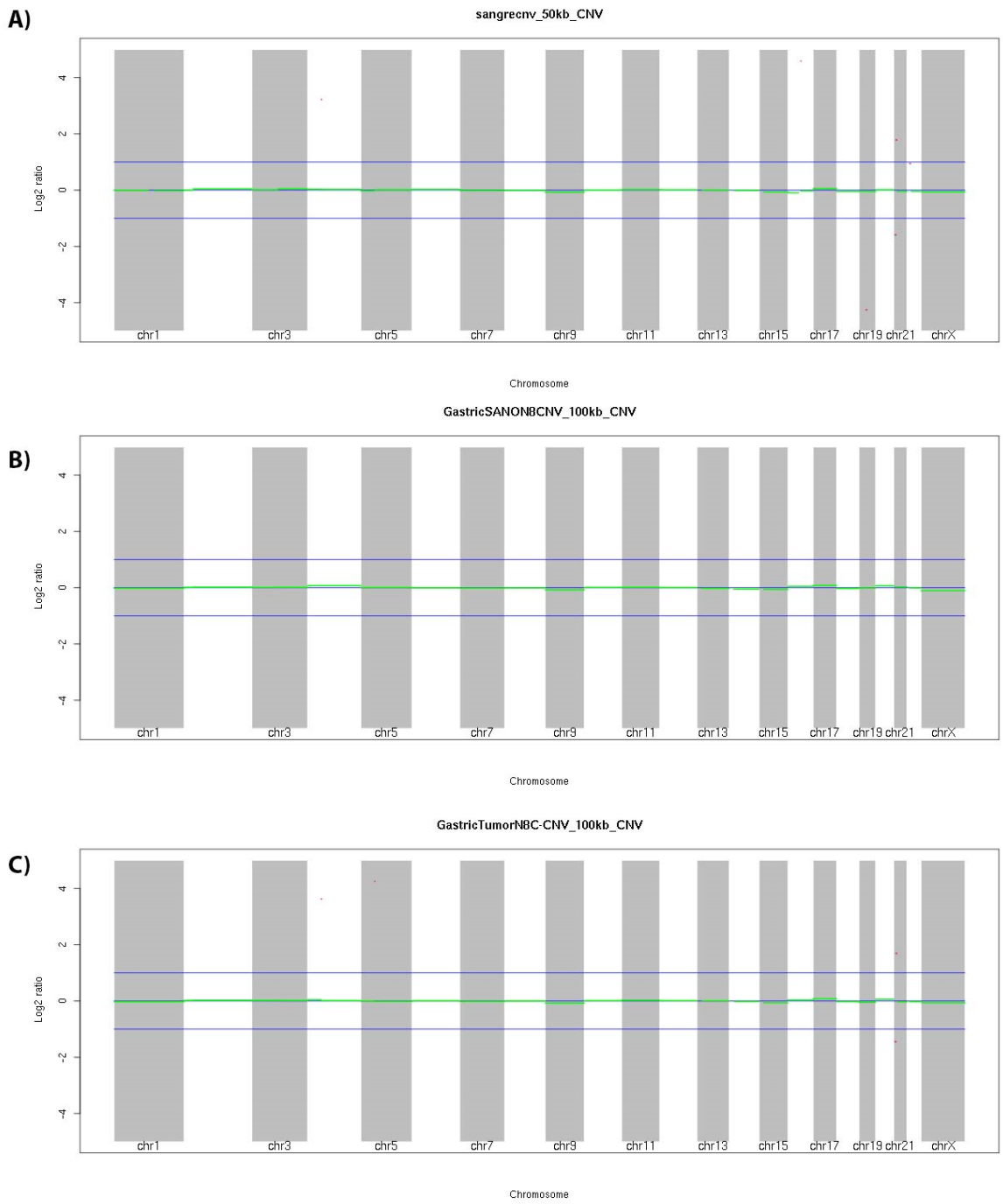


Figura 31. Los tejidos del paciente 8 PMBC, tejido sano y tumor no presentan variación en el número de copias. Las líneas verdes representan el largo de cada cromosoma que están representados en gris con el nombre correspondiente y en blanco el inmediato. Se observa que todas la líneas caen cercanas al cero en el rango para Log2CopyRatio (-0.3,3) por lo tanto no existe CNV. Además, los puntos rojos que se interpretan como ganancia de material genético (ver sección HiNT-CNV) en este caso corresponden a regiones centroméricas no informativas (ver regiones rojas compartidas entre el cromosoma 21 de las PMBCs “Sangre” y tumor).

Análisis de translocaciones y CNVs en la línea celular SNU16 y el tumor T2000877

Esta tesis tiene como objetivo la estandarización de los programas que permitan la detección de variaciones estructurales en muestras de cáncer gástrico que seguirán ingresando al laboratorio y, en el caso de la muestra correspondiente al paciente 8 con cáncer gástrico del tipo difuso no se detectaron translocaciones ni variaciones en el número de copias. Adicionalmente se corrieron los flujos de trabajo bioinformático de análisis utilizando los datos disponibles de Ooi et al., 2020 correspondientes a la línea celular de CG SNU16 y el tumor de CG del subtipo intestinal T2000877. Estos datos presentaron interacciones aberrantes enriquecidas en *trans* (Figura 32 A y B, ejemplos marcados con círculos azules).

En tanto a la variación en el número de copias detectada por HiNT-CNV se encontraron 57 en el tumor gástrico del tipo intestinal T2000877 de las cuales 12 representaron ganancia de material genético con un Log2CopyRatio(>=0.3), 16 son pérdida de material genético con un Log2CopyRatio(<=-0.3) y 29 se mantuvieron neutrales en el intervalo de Log2CopyRatio (-0.3,0.3). En el caso de SNU16 se reportaron 31 CNV de las cuales 20 fueron neutrales, 5 pérdidas de material y 6 ganancias de material genético (Figura 33 A).

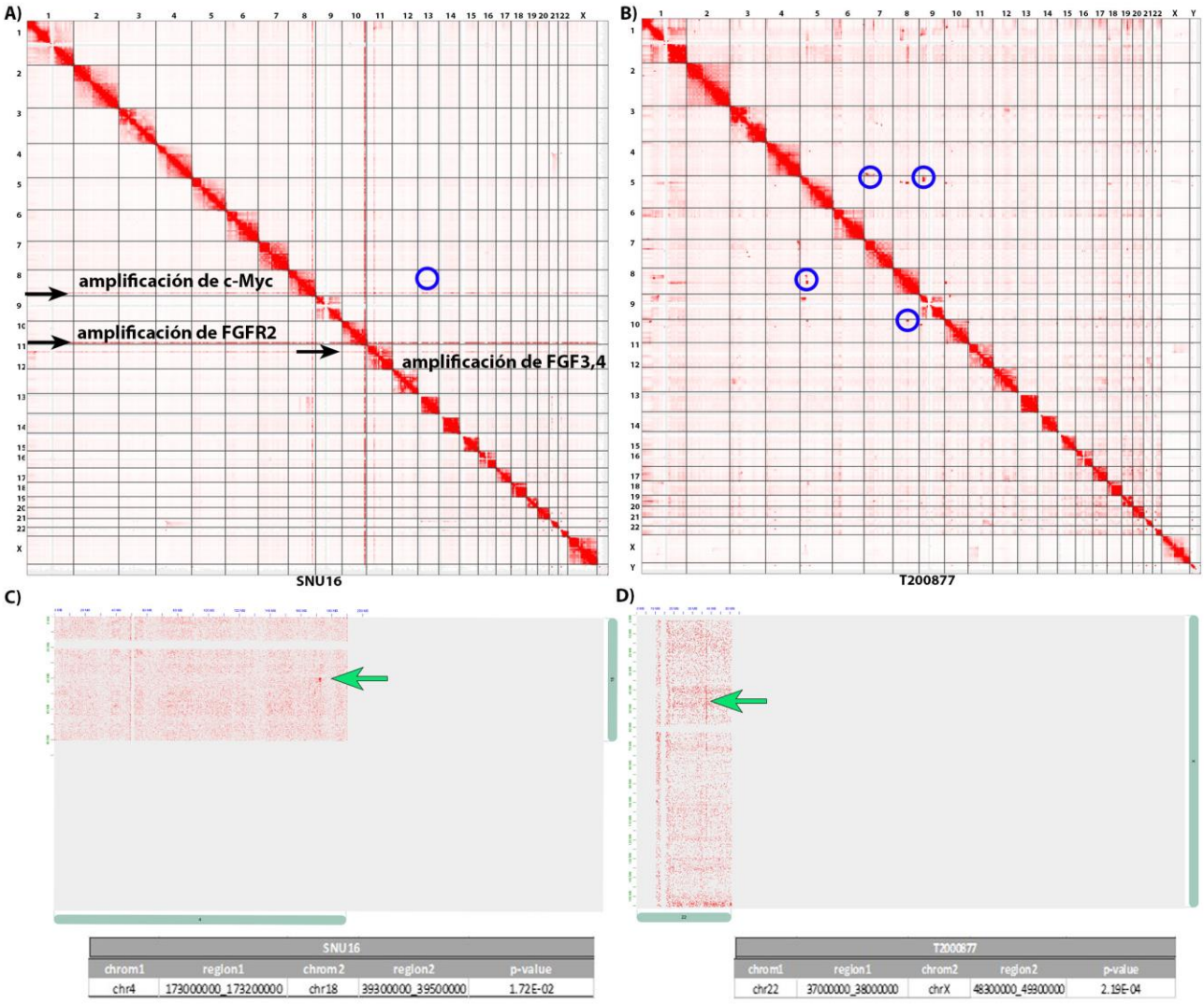
Respecto a las translocaciones encontradas con HiNT-TL fueron 54 en T2000877, a la resolución de un solo par de bases para el breakpoint se detectaron 10 translocaciones. En el caso de SNU16 se encontraron 21 y solamente en tres se detectó el breakpoint (Ver tabla 2 donde se muestran las coordenadas y breakpoints de estas translocaciones). Al corroborar estas coordenadas en las matrices e identificar las regiones en el 88enoma browser UCSC[87] para el caso de SNU16 las tres translocaciones detectadas a nivel de breakpoint corresponden a amplificaciones anteriormente reportadas[88] que se observan a lo largo de toda la matriz.

La amplificación del cromosoma 8 corresponde al gen c-Myc, el gen más comúnmente amplificado en cáncer[25]. En los casos de las amplificaciones de los

cromosomas 10 y 11 se trata de los genes del receptor 2 del factor de crecimiento de fibroblastos (FGFR2) y los factores de crecimiento de fibroblastos 3 y 4 (FGF3, FGF4), respectivamente (Figura 32 A matriz de SNU16, líneas paralelas que salen entre los cromosomas 8 “amplificación c-Myc”, 10 “amplificación FGFR2” y 11 “amplificación FGF3,4” y se comparten en todos los cromosomas en *trans*).

Este receptor y factores de crecimiento son fundamentales en el desarrollo embrionario, angiogénesis y en la tumorogénesis. En conjunto estas amplificaciones están relacionadas con el progreso del cáncer y la detección a escala de par de bases por medio de HiNT-TL demuestra la sensibilidad de esta herramienta al localizar amplificaciones genéticas.

En la figura 32 incisos C y D se muestran ejemplos de translocaciones que ocurrieron en estas muestras. En SNU16 la translocación de 2Mb entre el cromosoma 4 y 18 contiene el gen GALNTL6 que codifica para la N-acetylgalactosaminyltransterasa en la región del cromosoma 4 y cuyos polimorfismos se han asociado a mayor desempeño durante la actividad física y es fundamental en la señalización celular[89]. En el caso de T2000877 la translocación de 1Mb entre el cromosoma 22 y el X contiene en este último el gen de HDAC6, una desacetilasa de histonas cuyos inhibidores se utilizan como quimioterapia[90] (las regiones fueron mapeadas en el genoma browser UCSC)[87]. Estos ejemplos señalan la importancia del estudio de las variaciones estructurales en cáncer.



SNU16

chrom1	region1	chrom2	region2	p-value
chr4	173000000_173200000	chr18	39300000_39500000	1.72E-02

T200877

chrom1	region1	chrom2	region2	p-value
chr22	37000000_38000000	chrX	48300000_49300000	2.19E-04

Figura 32. Variaciones estructurales detectadas en la línea celular SNU16 y el tumor gástrico intestinal T200087. A y B) Matrices de Hi-C de SNU16 y T200087, se marcan en círculos azules algunos enriquecimientos de interacciones en *trans* característicos de las variaciones estructurales. **C y D)** Ejemplos de la visualización en las matrices de translocaciones (marcadas con flechas verdes) entre el cromosoma 4 y el 18 de 2Mb de longitud en el caso de SNU16 y el cromosoma 22 y el X con 1Mb en T2000877 (las coordenadas y la *p* asociada a cada una se muestran en las tablas inferiores).

Tabla 2. Lista de translocaciones encontradas por HiNT-TL en T2000877 y SNU16 para las cuales fue detectado el breakpoint a resolución de par de bases.

T2000877							
chrom1	region1	breakpoint1	chrom2	region2	breakpoint2	supportReads	p-value
chr7	52200000_53200000	52735468	chr19	30900000_31900000	31199778	2	0
chr5	32300000_33300000	32800858	chr8	70500000_71500000	70969555	1	0
chr5	26100000_27100000	26692744	chrX	62500000_63500000	63141840	1	0.0000937
chr22	37000000_38000000	37463797	chrX	48300000_49300000	48754596	1	0.000219
chr5	33700000_34700000	34142899	chr10	7300000_8300000	7679232	1	0.00539
chr3	98100000_99100000	98436136	chr7	91300000_92300000	91694330	1	0.0055
chr1	92600000_93600000	92921056	chr9	6200000_7200000	6710691	1	0.0125
chr1	50400000_51400000	50724870	chr10	6200000_7200000	6825499	1	0.0167
chr12	55300000_56300000	55988460	chr20	3800000_4800000	4293962	1	0.0266
chr10	21800000_22800000	22144734	chr14	102000000_103000000	102616365	1	0.0353
SNU16							
chrom1	region1	breakpoint1	chrom2	region2	breakpoint2	supportReads	p-value
chr8	127300000_128300000	127733371	chr10	120600000_121600000	121093168	1	0.00E+00
chr10	120100000_121100000	120912315	chr11	34900000_35900000	35192588	7	3.12E-05
chr8	126100000_127100000	126460499	chr11	34900000_35900000	35273467	1	1.25E-04

Cabe destacar que se detectaron en total 57 translocaciones en T2000877 y 31 en SNU16, por lo que la cantidad de lecturas quiméricas ambiguas que permitan la detección del breakpoint con esta resolución es baja (ver columna supportReads y sección anterior).

En el tumor de cáncer gástrico del tipo intestinal T2000877 se buscó la duplicación en tándem del *locus* que contiene al gen *CCNE1* que como ya se ha mencionado, su sobreexpresión derivada del secuestro de un potenciador adyacente al duplicarse la región, está relacionada con una pobre sobrevida en pacientes con cáncer gástrico[7]. Se utilizó como control la línea celular SNU16 que como ya se mostró es genómicamente más estable es decir se encontraron menos CNV y translocaciones, además de no encontrar alteraciones en el locus analizado.

Se pudo detectar la duplicación en tándem reportada por el grupo de Ooi, como se muestra en la matriz central de la figura 33 donde se compara en diagonal la matriz de SNU16 y la matriz de T2000877. Se observan dos regiones enriquecidas en interacciones en la región de la matriz correspondiente a T2000877 marcadas con círculos verdes.

La región en el cromosoma 19 que va del par de bases 29800000 al 30799999 con un Log2CopyRatio 2.511 y una p de 1.52135e-39 está marcada con línea azul en la

matriz B y un círculo del mismo color en el diagrama A CNV correspondiente a T2000877 mostrando que HiNT detectó una ganancia de material genético debido a una mayor cobertura de la esperada, poniendo de manifiesto la capacidad exploratoria de este pipeline.

Además, en la matriz se marca con una línea naranja la región centromérica de este cromosoma donde el experimento de Hi-C no detecta interacciones. A pesar de esto, en el diagrama A de CNV en naranja se marca en un círculo la supuesta variación en el número de copias encontrada por HiNT-CNV pero como ya se mencionó esta corresponde una región centromérica, este fenómeno se observa anteriormente en los datos del paciente 8 como ya se discutió y puede estar relacionado con errores de mapeo en regiones repetidas.

Por último, estos datos fueron analizados con HiCexplorer[74] y se obtuvieron los TADs una vez normalizadas por lecturas las matrices para hacerlas comparables (estos resultados serán mostrados a detalle en la siguiente sección). Se graficaron los TADs presentes en esta región y se encontraron diferencias derivadas de esta VE (Ver figura 33, C).

La recapitulación de los datos encontrados por el grupo de Ooi en el 2020 con estas muestras permite probar la capacidad de detección de variaciones estructurales de HiNT más allá de las translocaciones, tomando como punto de partida las CNV y contrastando los datos de coordenadas con las matrices. Si bien la interpretación correcta de estas VE y su impacto en la regulación genética también depende de otros datos genómicos como la detección de potenciadores; Esta puede ser una herramienta poderosa para la identificación de regiones alteradas.

En conjunto, estos datos muestran que los tejidos del paciente 8 son genómicamente estables, una característica atribuida a algunos tumores de cáncer gástrico del tipo difuso[4,9]. Por otro lado, se encuentran CNV y translocaciones en la línea celular de cáncer gástrico SNU16 y el tumor de cáncer gástrico del tipo intestinal T2000877 cuyo subtipo se ha relacionado con inestabilidad genómica[4,9].

El análisis de ambos conjuntos de datos muestran que HiNT permite encontrar CNV y translocaciones de manera muy sensible si se toma en cuenta que las matrices de los datos del grupo de Ooi fueron generadas con menos lecturas útiles de Hi-C que las generadas con los tejidos del paciente 8 (Ver figura 27). Además, esta herramienta puede guiar la búsqueda de otras VE no solo translocaciones, como lo demuestra la recapitulación de la duplicación en tándem del locus que contiene al gen CCNE1 tomando en cuenta las coordenadas de CNV.

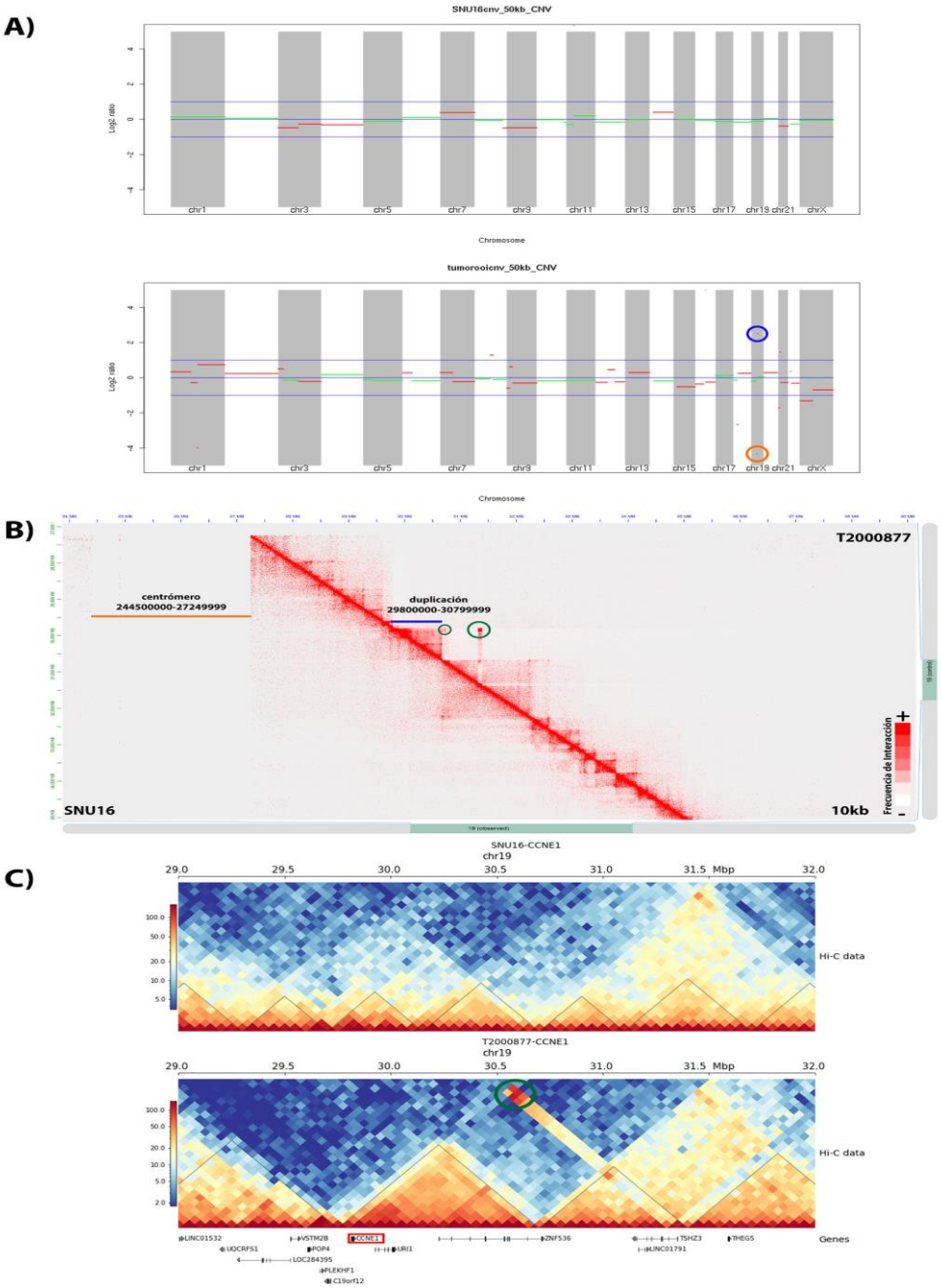


Figura 33. Recapitulación de duplicación en tándem del locus de CCNE1. A) Diagramas de CNV para T2000877 marcando en un círculo azul la correspondiente a la duplicación del locus que contiene a *CCNE1* y en naranja la correspondiente al centrómero. En SNU16 se muestran la mayoría de CNV como neutrales en verde, con un rango de Log2CopyRatio (-0.3,0.3).**B)** Matrices de Hi-C correspondientes a SNU16 en la diagonal inferior y a T2000877 en la superior, se marcan con círculos verdes las interacciones *de novo* que se forman por la variación estructural, la línea azul marca el largo de la variación del número de copias detectada por HiNT-CNV y la naranja la detectada para el centrómero del cromosoma 19.**C)** TADs presentes en la región rearreglada, comparación entre T2000877 y SNU16 presentando perdida de fronteras en el tumor atribuido a la VE, además de nuevos contactos observados en las matrices, círculo verde (este plot está realizado con HiCexplorer con matrices normalizadas por número de lecturas).

Identificación de TADs en las bibliotecas de Hi-C analizadas

La organización tridimensional del genoma es afectada por patologías como el cáncer[44,63], como se ha discutido anteriormente y se ha ejemplificado con los datos de Ooi *et al*. 2020 del tumor de cáncer gástrico del tipo intestinal T2000877 y la línea celular SNU16. Por lo tanto la identificación de TADs en los tejidos PMBCs, tejido adyacente sano y tumor de cáncer gástrico del subtipo difuso del paciente 8 puede guiar la búsqueda de regiones alteradas a nivel topológico en esta patología.

Las librerías de Hi-C SNU-16, T2000877 y PMBCs, tejido adyacente sano y tumor del paciente 8 fueron analizadas con el pipeline HiCexplorer[74] que permite la generación y normalización de matrices de Hi-C así como la identificación de TADs (Ver materiales y métodos).

En concreto en este análisis se normalizó la matriz de SNU16 respecto a T2000877, en tanto a las matrices del paciente 8 se normalizaron respecto al tejido sano (ambas matrices contienen la menor cantidad de lecturas válidas de Hi-C, ver sección de Reportes de HiCUP). Se corrigieron las matrices por cobertura, haciéndolas comparables entre sí y posteriormente se llamaron TADs a una resolución de 50kb (Ver materiales y métodos).

Se encontraron **3779** TADs en la línea celular de cáncer gástrico SNU16 y **4187** TADs en el tumor gástrico del subtipo intestinal T2000877. Respecto a los tejidos del paciente 8 se encontraron **3359** TADs en PMBCs, **3444** en el tejido adyacente sano y **3697** TADs en el tumor gástrico del tipo difuso (Figura 34 A). Siendo valores similares entre tejidos como era esperado, considerando que los TADs son estructuras conservadas entre tipos celulares[44].

Además, se llevó a cabo una comparación de los TADs compartidos entre el tejido adyacente sano y el tumor del paciente 8, ya que son los tejidos más parecidos en origen y los cambios en esta métrica pueden ser atribuidos al cáncer. Los resultados exponen que de los **3390** TADs comparados entre ambos tejidos, los mismos comparten **1723** TADs y **1667** TADs son exclusivos del tumor o diferenciales, es

decir las coordenadas de sus fronteras no se comparten con el tejido adyacente sano (Figura 34 B).

Esto representa un **49.17%** de los TADs presentes en el tumor de cáncer gástrico difuso del paciente 8. Por lo tanto, el tumor presenta un cambio topológico global que puede ser importante al desregular transcripcionalmente los genes contenidos en dichos dominios, por lo que un análisis detallado en estos *loci* será muy relevante, especialmente en este escenario en donde no existe una inestabilidad genómica muy acentuada.

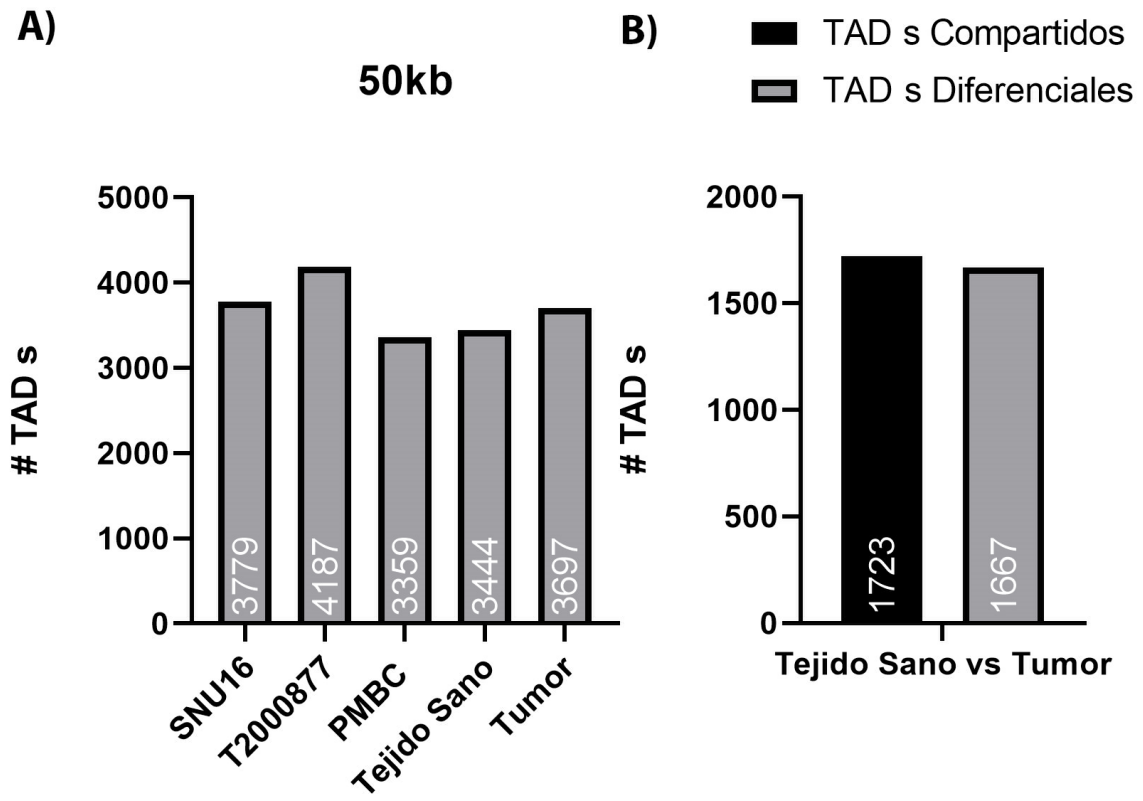


Figura 34. El numero de TADs encontrados por HiCexplorer en cada biblioteca analizada es equivalente. Cerca de la mitad de los TADs del tumor difuso del paciente 8 son exclusivos del mismo. **A)** Número de TADs encontrados en los datos de Ooi *et al* 2020 y los tejidos del paciente 8. **B)** TADs compartidos y diferenciales entre el tejido adyacente sano y el tumor del paciente 8. .

Análisis de expresión en células individuales en muestras del tejido adyacente sano y el tumor de CG del tipo difuso del paciente 8

El proyecto del que forma parte esta tesis considera el procesamiento de las muestras de los tumores de cáncer gástrico y el tejido adyacente sano de cada paciente mediante Hi-C y scRNA-seq, buscando obtener la información topológica y de expresión genética a nivel de células individuales en cada muestra para poder integrar la información topológica con la regulación de la expresión de genes en la misma muestra.

En concreto, en el caso del paciente 8 estos tejidos fueron los primeros en el proyecto relacionados con cáncer gástrico procesados mediante scRNA-seq, además de las bibliotecas de Hi-C obtenidas que se han analizado a lo largo del texto. Las suspensiones de células únicas de las muestras fueron procesadas mediante el protocolo de aislamiento basado en Droplets 10XChromium[64] que utiliza un chip de microfluídos. La estrategia experimental se describe a detalle en la sección de materiales y métodos.

Expresión de células individuales en el tejido adyacente sano y tumor gástrico difuso del paciente 8

Se analizaron los datos del tejido sano adyacente y el tumor gástrico difuso del paciente 8 con el pipeline de 10XCellRanger que realiza el alineamiento y una descripción de calidad de los datos como se muestra en la tabla 3 (Ver materiales y métodos). En el tejido adyacente sano se obtuvo un número celular, un número de lecturas y de genes promedio por célula esperado, con un 88.6% de alineamiento al genoma de referencia hg38. Sin embargo, en el caso del tumor solo se detectaron 277 células y 275 genes promedio por célula con un alineamiento al genoma de 65.3%.

Estos resultados pueden deberse a que durante el protocolo se consiguió una emulsión parcial en el experimento lo cual puede llevar al etiquetado y análisis de moléculas flotando en la muestra en lugar de dentro de células individuales.

Por otro lado este resultado puede relacionarse con la baja viabilidad inherente a las muestras tumorales y la presencia de ARN libre en la disolución. Probablemente afectando la formación de gemas y obteniendo algunas que no representan material de una sola célula.

A pesar de esto, con la exigencia de los controles de calidad que se mostrarán a continuación se espera un filtrado de datos estricto que arroje información relevante de esta muestra. Sin embargo, no se deja de lado el hecho de que estos resultados son preliminares y hace falta procesar más muestras para poder profundizar en las características transcriptómicas de este tipo de tumores.

Tabla 3. Datos de calidad de los experimentos de scRNA-seq obtenidos mediante 10XCellRanger en el tejido adyacente sano y tumor gástrico difuso del paciente 8.

Tejido	# Celular Detectado	Lecturas promedio por célula	Genes promedio por célula	Alineamiento al genoma
Tejido adyacente sano	2,748	84,927	1,164	88.6%
Tumor	277	380,498	275	65.3%

A partir de los archivos generados por 10XCellRanger **Features**, **Barcodes** y **Matrix** (Ver materiales y métodos) se continuó con el análisis de estos datos mediante la paquetería de R Seurat que permite identificar células únicas y vivas según el siguiente estándar, eliminando los datos que contienen menos de 200 genes detectados y se mantienen los que contienen arriba de 2500 genes con un porcentaje menor al 5% de expresión de genes mitocondriales.

Esto parte del racional de disminuir el error de procesar datos de perlas que solo contienen transcritos flotantes y eliminar las células muertas en la muestra. Todos los resultados mostrados de manera subsecuente fueron analizados con Seurat, en la sección de materiales y métodos se describe su funcionamiento.

Para realizar este control, con los datos del paciente 8 se graficaron los diagramas de violín que consideran los genes detectados, el número de lecturas para cada gen y el porcentaje de genes mitocondriales para las muestras sano y tumor (Figura 35

A y B respectivamente) se decidió mantener el estándar de 5% de genes mitocondriales en el tejido sano y un 20% en el tumor. El porcentaje de genes mitocondriales se muestra en la tercera gráfica de violín de la figura 35 para ambos tejidos.

Una vez realizado este control de calidad obtuvimos **1209** células en el tejido adyacente sano con un total de genes detectados de **18535** y para el tumor **146** células con **8501** genes detectados. Los experimentos de scRNA-seq en cáncer gástrico publicados son la suma de varios tejidos tumorales, procurando la diversidad celular, este es una primera estandarización para seguir analizando tumores de CG por esta metodología.

Cabe mencionar que en el caso del tejido tumoral, experimentalmente se obtuvo una emulsión parcial durante el protocolo lo cual explica el bajo número celular detectado, como se menciona anteriormente. A pesar de esto, el número de células y genes encontrados una vez realizados los controles de calidad de forma estricta eligiendo 20% de genes mitocondriales en lugar de 30% como se estila en trabajos con tumores y en particular con tumores gástricos[65,66,78,79], pretende poder realizar comparaciones informativas entre los dos tejidos.

Estos datos comprenden células únicas y vivas debido a su complejidad transcripcional y porcentaje de expresión de genes mitocondriales, por lo tanto se continuó con su análisis.

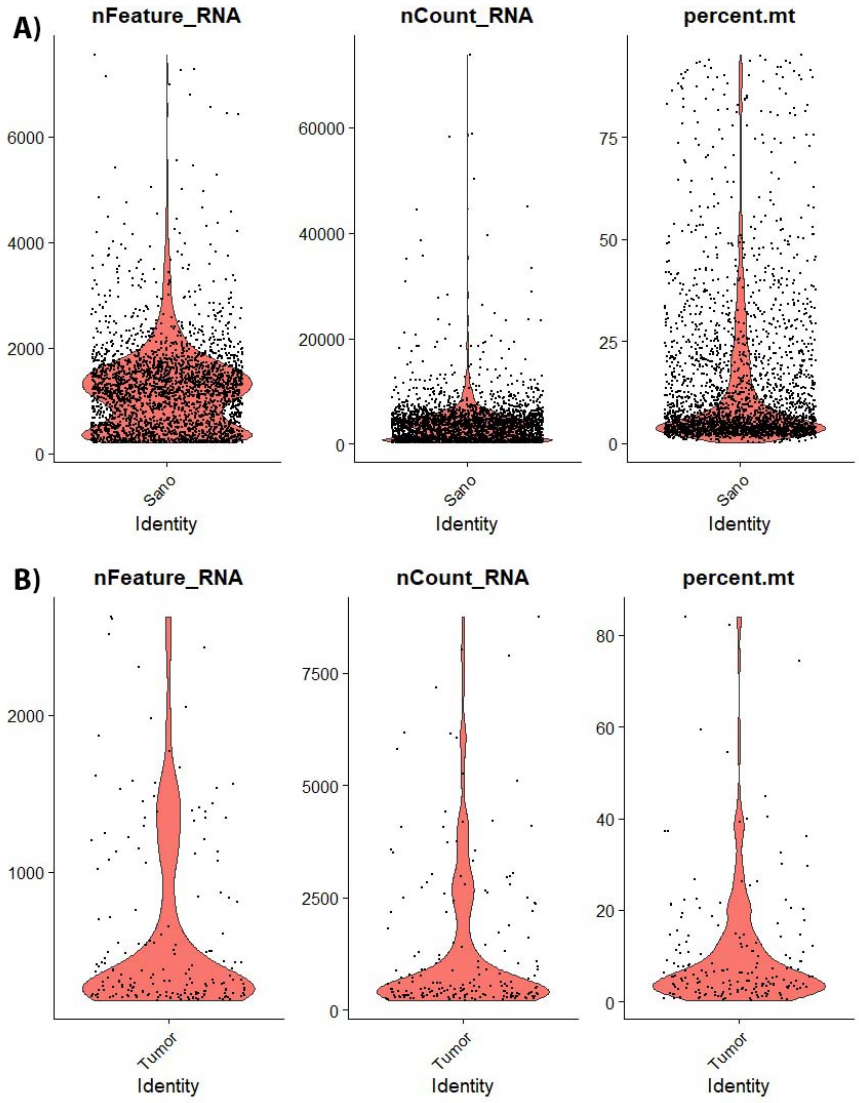


Figura 35. Gráficos de violín para el tejido adyacente sano (A) y tumor (B). nFeature_RNA representa la distribución de todos los genes detectados para cada tejido. nCount_RNA la distribución del número de lecturas por gen o "Feature" y percent.mt el porcentaje de genes mitocondriales detectados. Los análisis se realizan tomando en cuenta las regiones con mayor cantidad de datos en la distribución de nFeature_RNA y percent.mt.

Posteriormente, para ajustar la variabilidad en la expresión de los genes y poder comparar las muestras, se buscan los 15 genes más variables (Figura 36 A tejido adyacente sano y B tumor) y se escalan haciéndolos comparables entre sí.

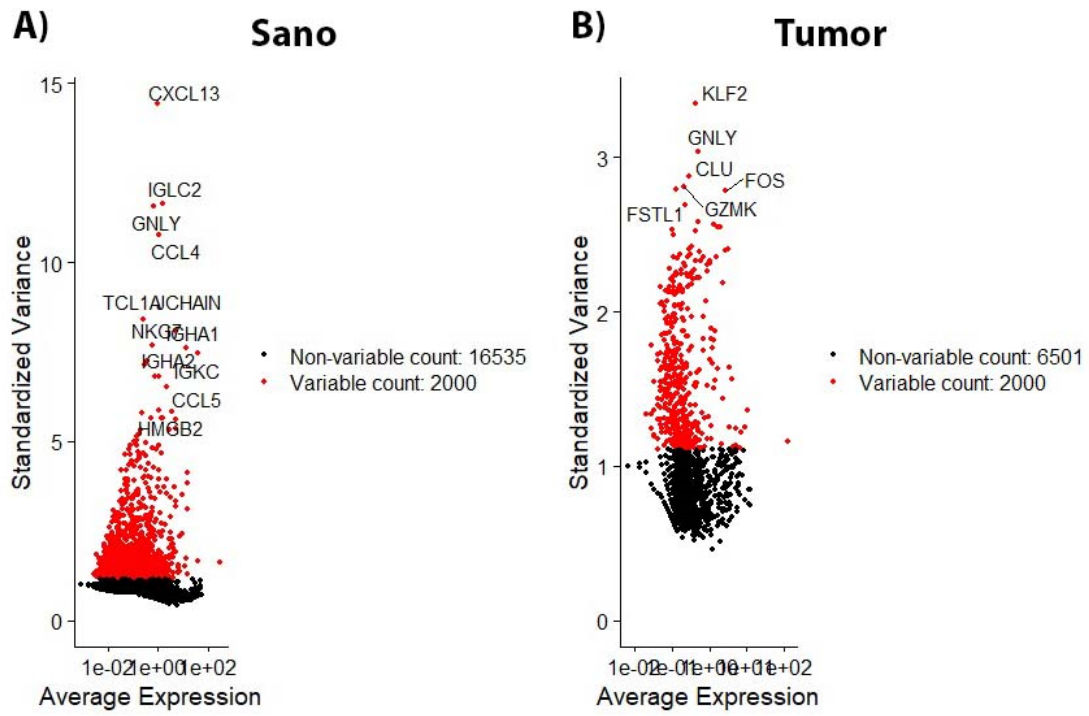


Figura 36. Genes con mayor varianza en cada uno de los tejidos a partir de los cuales se ajusta la expresión de los demás. El promedio de expresión y la varianza de todos los genes detectados se ajustan a una escala logarítmica (ejes *X* y *Y* de los gráficos) que permite la comparación de expresión entre genes.

Una vez ajustados los datos se realiza un análisis de componentes principales PCA para reducir las dimensiones de los mismos mediante la agrupación de los datos (en este caso células) por su varianza y se genera una gráfica denominada diagrama de codo que muestra el número de componentes principales PC que capturan mejor la varianza, se espera ver una meseta a partir del PC 10. Es decir, variables que ya no capturan más varianza. Este paso es fundamental para la reducción lineal de dimensiones (Ver sección de métodos) (Figura 37 A tejido adyacente sano y B tumor de CG difuso).

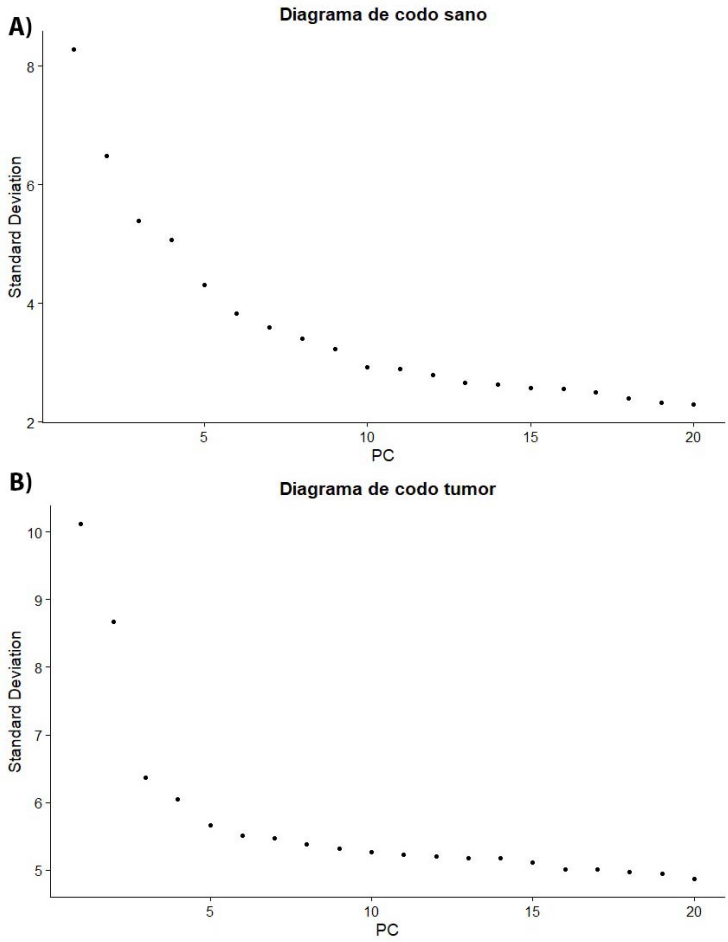


Figura 37. Diagrama de codo para cada tejido. Se observa la formación del codo en el PC 10, si se considera para el análisis un número mayor a 10 PC ya no se captura mayor varianza. Por lo tanto, el número de PCs que se usaron en los análisis posteriores fueron 10.

A partir del PCA se puede realizar una representación gráfica de clusters definidos por su expresión genética y relacionar estos perfiles de expresión con bases de datos de tejidos, en este caso se utilizó la base HumanPrimaryCellAtlas y las identidades se representaron mediante el algoritmo UMAP (Ver sección de métodos donde se describe con detalle el algoritmo).

En el tejido sano se encontraron tres identidades: estómago, sangre y nódulo linfático. En el caso de tumor solo se encontró estómago y nódulo linfático. Además, se obtuvieron las listas de genes que definen a cada tipo celular (Figura 38).

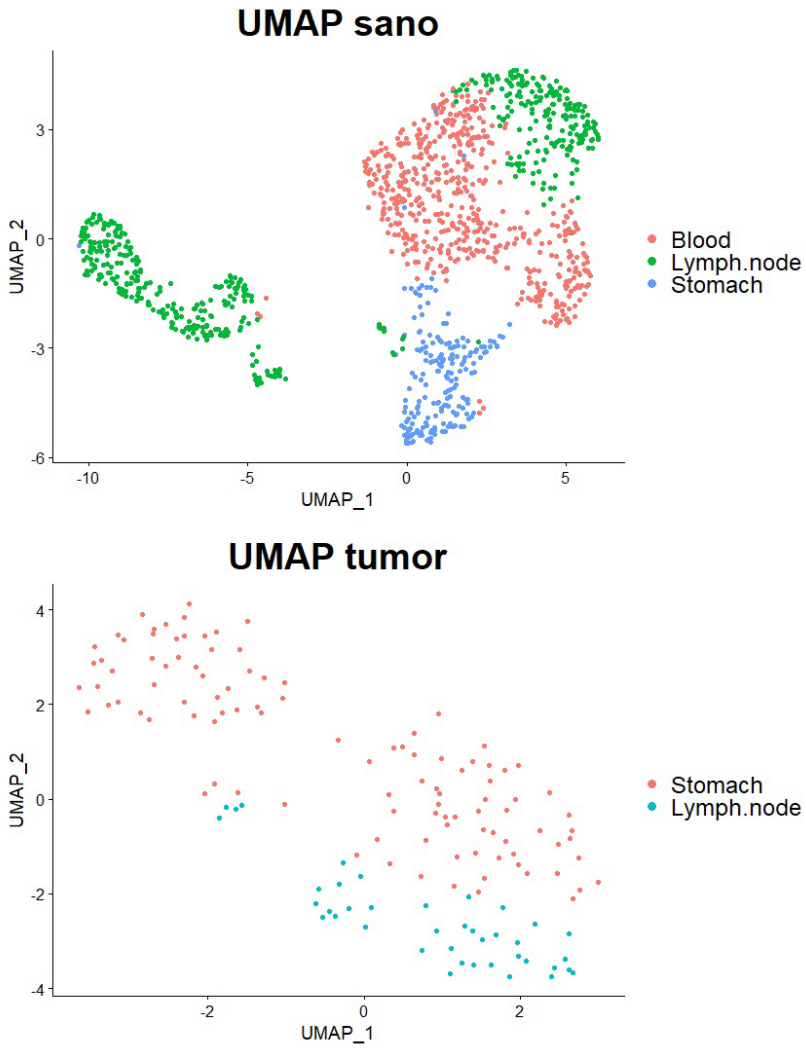


Figura 38. Las muestras recuperan identidades celulares de tejidos gástricos. UMAP de las muestras sano y tumor. Podemos observar que ambas muestras contienen los tejido estómago y nódulo linfático tomando en cuenta la base de datos HumanPrimaryCellAtlas.

Este resultado es esperado ya que se recuperan identidades celulares presentes en el tejido gástrico. Cabe mencionar que la falta de la identidad “sangre” en el tumor corresponde con la biología no vascularizada de los tumores en etapas iniciales. Sin embargo, también puede corresponder con el bajo numero celular recuperado en el experimento. Indagar más en la diversidad celular en estos tejidos y corroborar estas observaciones dependerá de los resultados obtenidos en muestras futuras.

Sin embargo, a pesar de las pocas células se recuperan en el tumor, las identidades celulares encontradas son pertenecientes a tejido estomacal, esto muestra que la complejidad obtenida es suficiente para los análisis de identidad y robustece los análisis posteriores.

Los UMAP y los genes que considera para cada cluster se analizaron como tumor y sano mediante MAST (Figura 39) (Ver sección de métodos) permitiendo analizar los genes diferencialmente expresados entre cada muestra encontrando **1811** genes diferencialmente expresados, de los cuales **1325** están sobreexpresados y **486** subexpresados en el tumor respecto al tejido sano.

El uso de esta estrategia corresponde con el hecho de que en los experimentos de Hi-C no es posible distinguir entre tipos celulares. Realizando comparaciones entre la expresión del tejido adyacente sano y el tumor de CG del tipo difuso del paciente 8 en general sin considerar la diversidad celular de los mismos, estos datos se pueden contrastar con los datos topológicos de Hi-C.

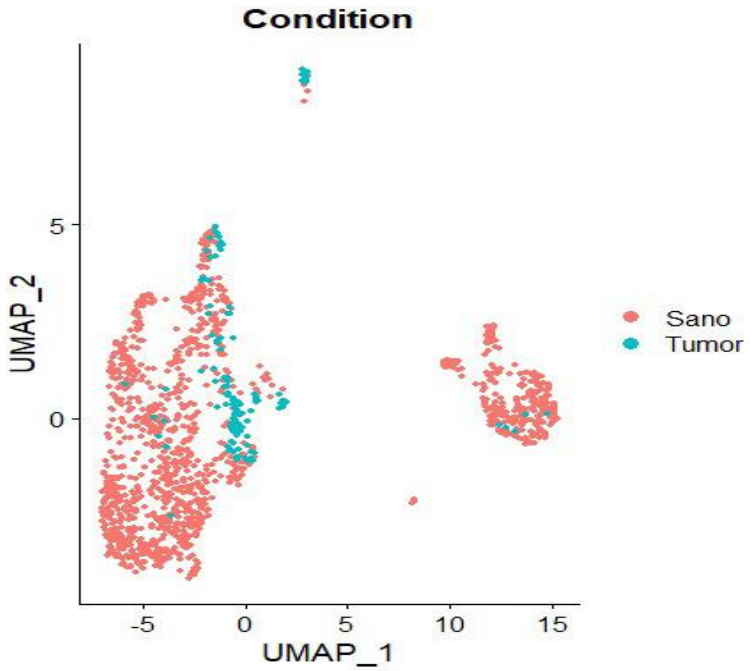


Figura 39. UMAP del tejido sano y el tumor. Se muestran las similitudes de varianza entre el tejido adyacente sano y el tumor, cabe destacar que se obtienen 3 grupos con células representantes de cada tejido, siendo consecuente con las identidades celulares previamente detectadas (Ver figura 38).

Finalmente, los 1325 genes sobreexpresados en el tumor gastrico difuso fueron analizados con la herramienta web Enrichr[82–84], que analiza listas de genes mediante diversas herramientas arrojando categorías ontológicas tales como procesos biológicos en dónde están involucrados y búsqueda de factores de transcripción relacionados con los GDE (Ver materiales y métodos).

Específicamente para los fines de este trabajo se utilizó la base de datos REACTOME[85,86] donde de encontró que la via más enriquecida según los genes expresados a la alta en el tumor de CG difuso del paciente 8 es la carga de cohesina a la cromatina con un score combinado de 476.20 (Ver materiales y métodos). Esta métrica considera la importancia de los genes relacionados a la vía y la p ajustada de la expresión de los mismos (tabla 4). Concretamente, los genes sobrexpresados son las subunidades del complejo de cohesina *SMC1A*, *SMC3*, *STAG2* y *RAD21* y las proteínas de carga y mantenimiento del complejo de cohesina a la cromatina *WAPL*, *NIPBL* y *PDS5A*[92,93].

Tabla 4. Vías relacionadas con los genes sobreexpresados en el tumor de cáncer gástrico difuso respecto al tejido adyacente sano.

Índice	Nombre	Valor de p	p-ajustada	Razón de probabilidades	Puntuación Combinada
1	Cargado de la cohesina a la cromatina R-HSA-2470946	5.542e-7	0.00001605	33.06	476.20
2	ARNm Splicing R-HSA-72172	1.748e-21	6.076e-19	5.99	286.40
3	ARNm Splicing-Vía mayor R-HSA-72163	6.913e-21	1.922e-18	6.04	280.26
4	Generación de moléculas de segundos mensajeros R-HSA-202433	3.721e-10	2.351e-8	12.57	272.86
5	Prosesamiento de Intrones con CAP PreARNm R-HSA-72203	1.383e-21	6.076e-19	5.16	247.75
6	Regulación de la respuesta inmune y migración por RUNX3 R-HSA-8949275	0.0002582	0.002300	28.27	233.57
7	Sistema inmune R-HSA-168256	5.112e-37	7.106e-34	2.70	225.50
8	Telofase/Citokinesis Mitotica R-HSA-68884	0.000006636	0.0001264	16.53	197.03
9	Establesimiento de la cohesión de las cromátidas hermanas R-HSA-2468052	0.00002891	0.0004143	16.99	177.52
10	Interacción de RUNX1 con cofactores cuyo efecto en los blancos de RUNX1 es desconocidoR-HSA-8939243	1.817e-8	7.654e-7	9.49	169.07

Se hace enfásis a la carga de la cohesina a la cromatina por su importancia en la organización tridimensional del genoma.

Este resultado es importante ya que como se menciona en la introducción, la cohesina es la encargada de llevar acabo el proceso de extrusión de loops de cromatina junto con otras proteínas estructurales[15,92,93] que ponen en contacto los locus genómicos y permiten la regulación adecuada de la expresión de los genes.

Por lo tanto, la alteración de esta vía en el tumor gástrico difuso del paciente 8 puede ser relevante en la alteración de la topología del genoma de las células tumorales y coincide con los cambios descritos a nivel de TADs en el tumor en comparación al tejido adyacente sano. Esta observación resulta muy interesante ya que propone un posible mecanismo de alteración topológica que conlleve a la desregulación transcipcional fina en este tipo tumoral con pocas alteraciones genéticas. Conocer si esto es una característica general de este subtipo tumoral requerirá de más estudios incluyendo un mayor número de muestras

Integración de datos topologicos y trancripcionales en CG difuso

Finalmente, mediante la herramienta de visualización de datos genómicos Seqmonk se mapearon las coordenadas de los 1667 TADs exclusivos del tumor de CG difuso del paciente 8, además de las coordenadas de los 1811 GDE en el mismo y se realizó un analisis para conocer que GDE se encontraban dentro de los TADs exclusivos del tumor. Cabe mencionar que solo se mapearon 1754 GDE ya que 57 transcritos mapeaban a regiones que no contenía la versión del genoma de referencia utilizada (GRCh38_v108). Sin embargo, todos correspondian a coordenadas redundantes de transcritos ya considerados en los 1754 GDE analizados. Se encontró que de los **1667** TADs exclusivos del tumor **452** TADs contienen **717** GDE de los cuales **537** genes se encuentran sobreexpresados y **180** genes subexpresados en el tumor (Ver Figura 40).

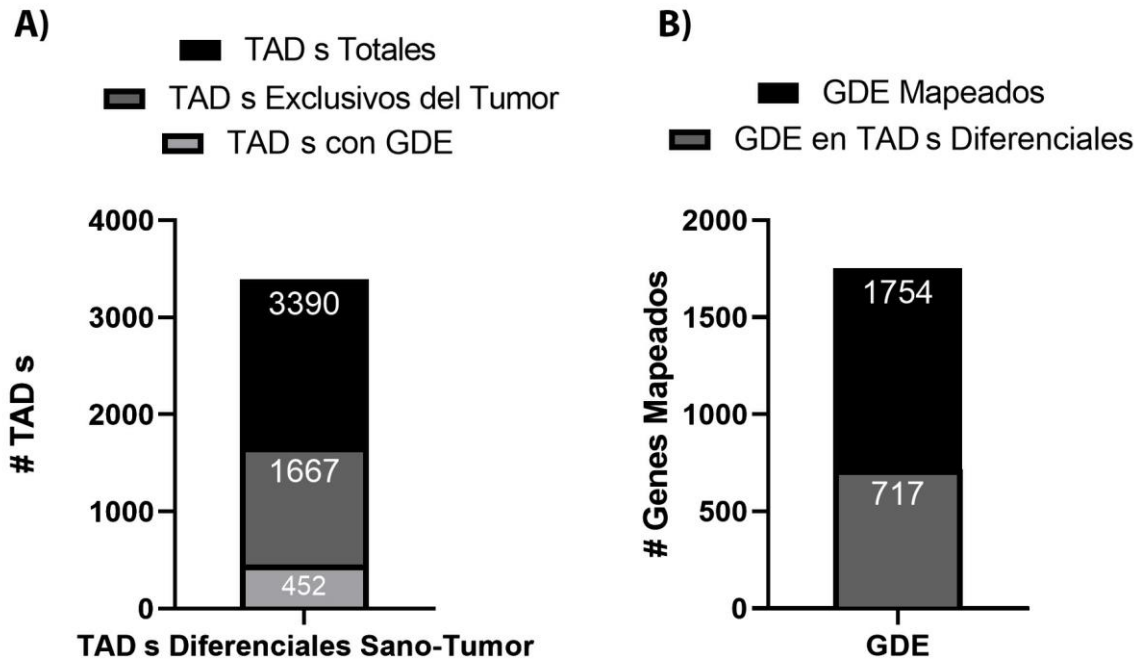


Figura 40. Número de genes alterados contenidos en TADs exclusivos del tumor. Se encontró que 452 TADs exclusivos del tumor contienen 717 GDE. **A)** Gráfico de barras de TADs exclusivos del tumor (1667) y TADs que contienen GDE (452). **B)** Gráfico de barras del número de GDE mapeados en Seqmonk (1754) y GDE que se encuentran en TADs diferenciales (717) o exclusivos del tumor.

Debido a su importancia en la topología del genoma se decidió indagar en los *loci* donde se encuentran los genes que forman parte de la cohesina y que están involucrados en su asociación con la cromatina. Esto se realizó graficando las matrices de contacto de estas regiones mediante HiCexplorer agregando los TADs detectados por el pipeline y las coordenadas de los genes presentes en las mismas, considerando su estado transcripcional (Ver materiales y métodos).

Esto es informativo ya que muestra evidencia del impacto de los cambios topológicos en la regulación genética de estos *locus*. Sin embargo, no se profundizó en la función biológica de los genes que comparten locus con los genes estructurales y regulatorios de la cohesina, solo se hace enfásis en su estado transcripcional, por las razones mencionadas anteriormente.

Los genes que codifican a las subunidades energéticas de la cohesina *SMC1A* y *SMC3* se encuentran en TADs exclusivos del tumor

En el tumor de CG difuso del paciente 8 las subunidades de la cohesina involucradas en la actividad ATP dependiente de la misma SMC1A y SMC3[28–30] se encuentran en TADs exclusivos del tumor (Ver Figura 41 área sombreada).

En los locus donde mapea cada gen se observa un proceso equivalente donde se generan dos TADs *de novo* en el tumor a partir de uno solo presente en el tejido sano. Así como un cambio general de los TADs encontrados en la región de 4Mbp considerada en este análisis entre tejidos (Ver Figura 41 flechas rojas).

En el caso del TAD donde se encuentra *SMC1A* se encontró sobreexpresado también el gen *HUWE1* (rojo) y subexpresado el gen *TSPY2* (azul). *SMC3* está sobreexpresado unicamente en el TAD que lo contiene. Cabe señalar la sobreexpresión general de los genes contiguos al TAD adyacente *PDCD4*, *ADRA2A* y *SHOC2* contenidos en otro TAD *de novo* exclusivo del tumor.

Estas observaciones apoyan la hipótesis de que el cambio global en la topología genómica está relacionado con el cambio en la expresión genética de las células.

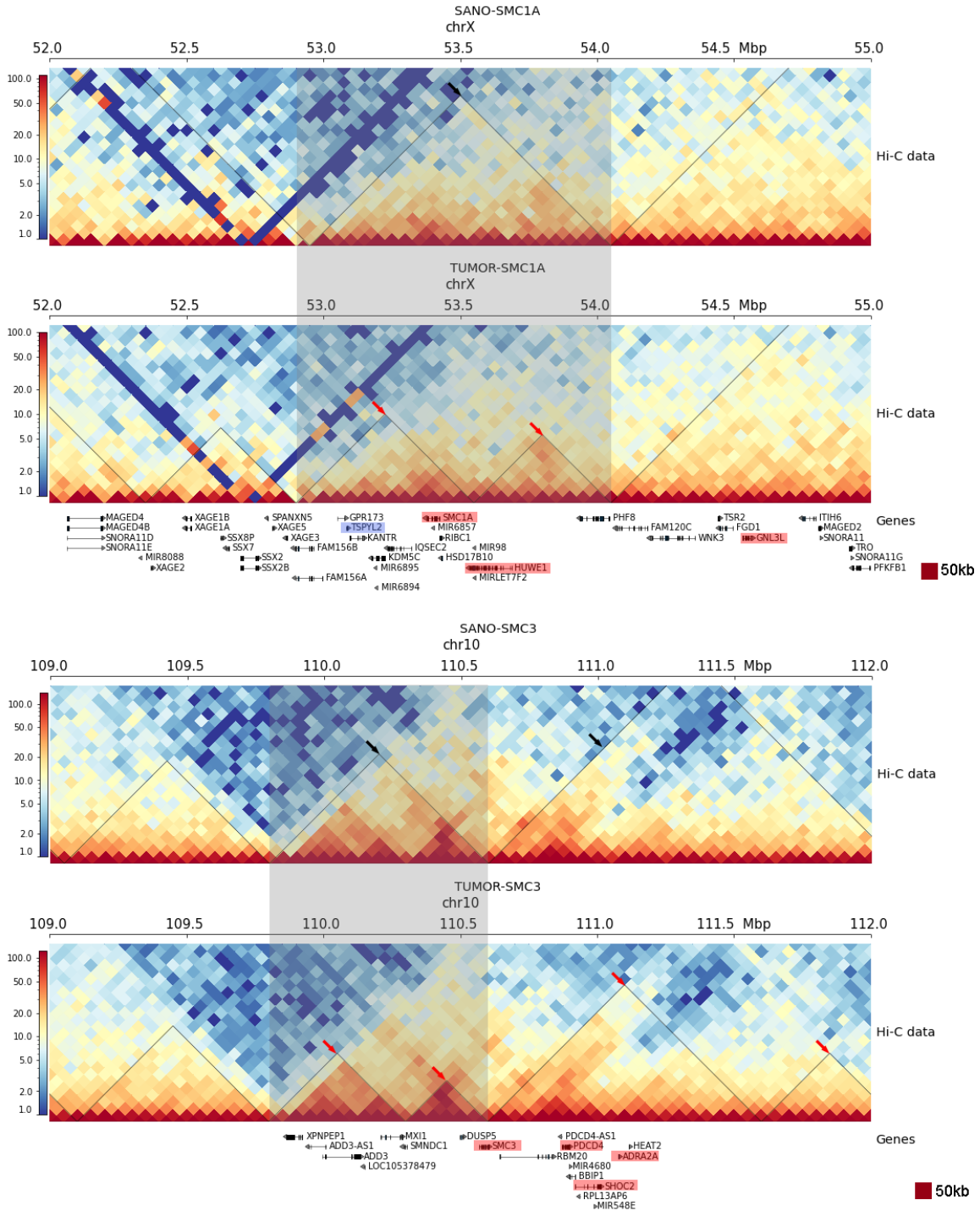


Figura 41. Los genes de las subunidades de la cohesina *SMC1A* y *SM3* se encuentran en TADs exclusivos del tumor de CG difuso del paciente 8 (área sombreada). Comparación de matrices de Hi-C entre tejido adyacente sano y tumor del locus que contiene los genes *SMC1A* (arriba) y *SMC3* (abajo). Flechas negras TAD presente en tejido sano, Flechas rojas TADs *de novo* en el tumor. Genes sobreexpresados en rojo, genes subexpresados en azul (Matrices a 50kb de resolución).

Los genes de las subunidades estructurales de la cohesina *RAD21* y *STAG2* se encuentran en TADs exclusivos del tumor

RAD21 interacciona con SMC1A y SMC3 cerrando el anillo de la cohesina[28–30]. Mientras que STAG2 es fundamental para la asociación del complejo de cohesina con el ADN formando parte del núcleo del complejo de la cohesina. Ambos genes se encuentran en TADs diferenciales del tumor respecto al tejido sano (Ver figura 42 área sombreada).

En el caso de *RAD21* se localiza en un TAD *de novo* en el tumor formado por una nueva frontera que divide en dos al TAD presente en el tejido sano, siendo esta la única alteración topológica relacionada con los TADs en la ventana de 6Mbp graficada (Ver figura 42 arriba).

STAG2 se ubica en una región con un paisaje topológico ampliamente distinto respecto al tejido sano, considerando las 4Mbp ploteadas, en un TAD exclusivo (Ver figura 42 abajo).

Es importante recalcar que tanto *RAD21* como *STAG2* son los únicos genes sobreexpresados en estos *loci*.

Estos dos casos son contrastantes en el hecho de que si bien el cambio topológico es acompañado del cambio transcripcional en estos genes de interés. En particular, en la región de *STAG2* a pesar de los TADs *de novo*, no se reportan más cambios transcripcionales, ejemplificando la complejidad de la regulación genética fina y la diversidad de fenómenos particulares.

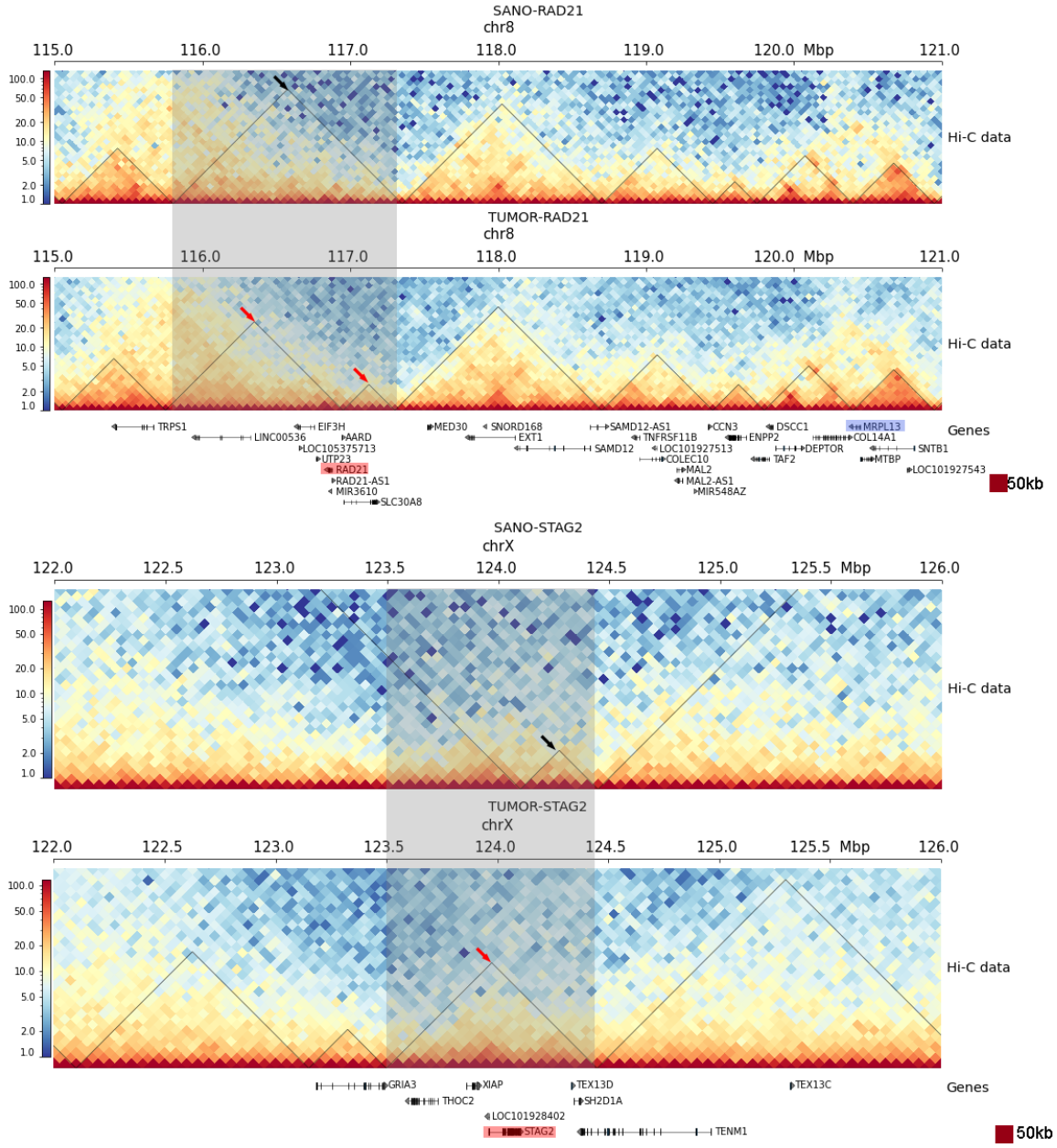


Figura 42. Los genes de las subunidades de la cohesina *RAD21* y *STAG2* se encuentran en TAD´s exclusivos del tumor de CG difuso del paciente 8. Además, son los únicos genes alterados transcripcionalmente en su TAD respectivo (área sombreada). Comparación de matrices de Hi-C entre tejido adyacente sano y tumor del locus que contiene los genes *RAD21* (arriba) y *STAG2* (abajo). Flechas negras TAD presente en tejido sano, Flechas rojas TADs de novo en el tumor. Genes sobreexpresados en rojo, genes subexpresados en azul (Matrices a 50kb de resolución).

Los genes de regulación del cargado de cohesina a la cromatina *WAPL* y *PDS5A* se encuentran en TADs exclusivos del tumor a excepción de *NIPBL*

Los genes que regulan el mantenimiento, cargado y descargado de la cohesina a la cromatina también están sobreexpresados en el tumor de CG del tipo difuso del paciente 8 respecto al tejido sano.

PDS5A que mantiene la asociación de la cohesina a la cromatina[28–30] se encuentra en un TAD exclusivo del tumor formado por un cambio en una de las fronteras del TAD que lo contenía originalmente en el tejido sano (Ver Figura 43 arriba, región sombreada). Además, el estado transcripcional sobreexpresado lo comparte con el gen *SMIM14*. De manera contraria, el gen *RPL9* está subexpresado. Ambos presentes en el mismo dominio. Aunado a esto, el TAD contiguo a la derecha contiene a los genes *N4BP2* y *RHOH* también sobreexpresados en el tumor, aunque se comparte el mismo TAD con el tejido adyacente sano. Hacia el final de la región de 4Mbp graficada se observa un cambio en los TADs presentes en el tumor definido por la pérdida de fronteras en la misma en comparación con el tejido sano (Ver Figura 43 arriba, flechas negras en tejido sano y roja en tumor). Estos resultados hacen interesante este *locus* que contiene a *PDS5A* para estudios posteriores en modelos de CG.

WAPL es la proteína encargada de la disociación de la cohesina de la cromatina[28–30] y su gen se encuentra también en un TAD exclusivo del tumor, formado por dos fronteras generadas en un TAD más grande presente en el tejido sano (Ver figura 43 abajo, región sombreada). Cabe señalar que en este TAD exclusivo solo se encuentra el gen *WAPL*. Por otro lado, en el TAD contiguo a la derecha aúnque también es exclusivo del tumor no se detectó ningún gen alterado transcripcionalmente.

Finalmente, *NIPBL* que regula el cargado de la cohesina a la cromatina[28–30], aunque está sobreexpresado en el tumor no se localiza en un TAD exclusivo del tumor. Además, en la región ploteada de 3Mbp se observa la subexpresión del gen *IL7R* a la izquierda y la sobreexpresión del gen *WDR70* a la derecha, ambos en TADs

equivalentes al tejido sano (Ver Figura 44). Este locus muestra que el cambio en los TADs no es el único proceso involucrado en los cambios de expresión de los transcritos.

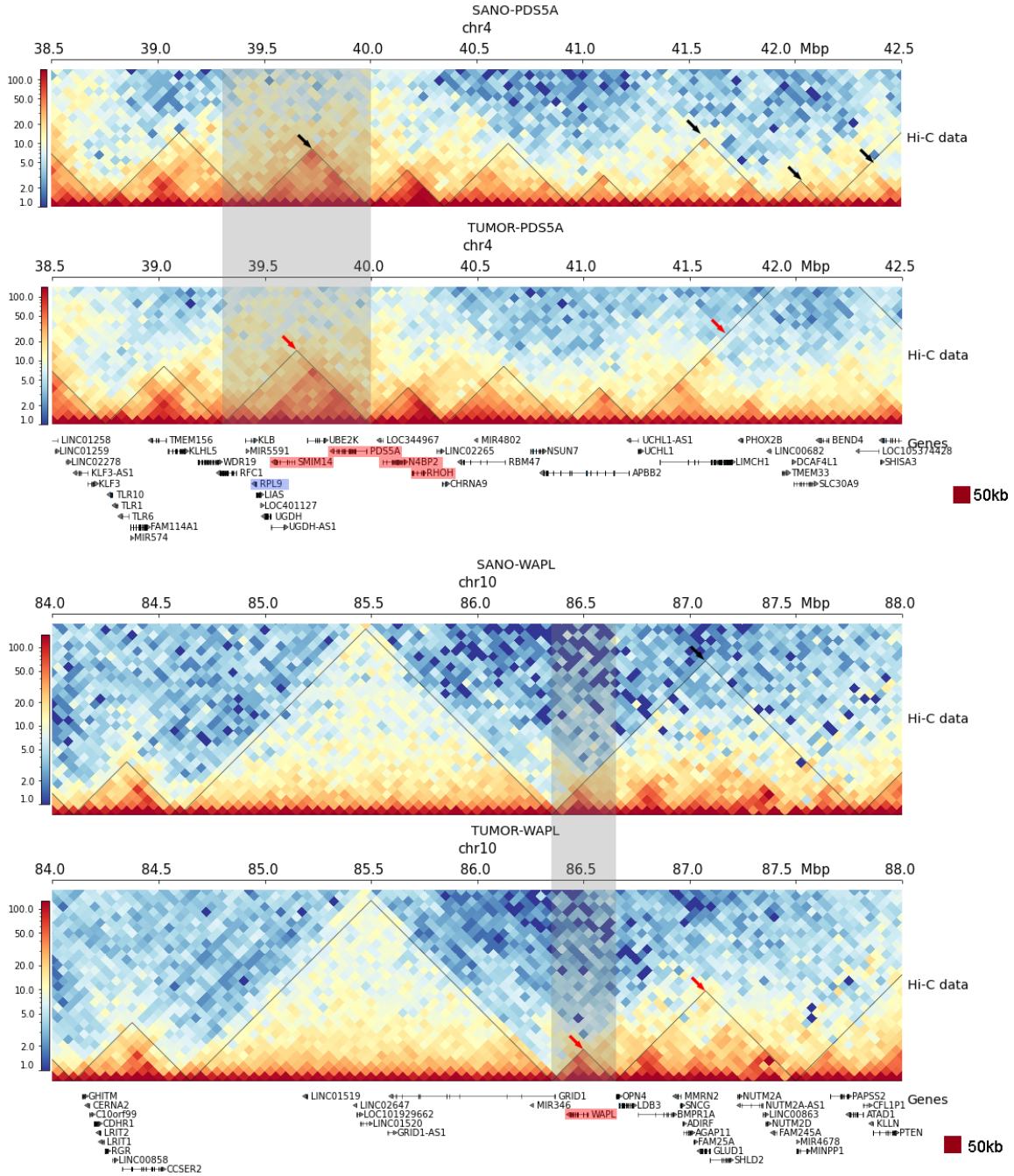


Figura 43. Los genes de mantenimiento *PDS5A* y descarga *WAPL* de la cohesina se encuentran en TADs exclusivos del tumor de CG difuso del paciente 8 (área sombreada). Comparación de matrices de Hi-C entre tejido adyacente sano y tumor del locus que contiene los genes *PDS5A* (arriba) y *WAPL* (abajo). Flechas negras TAD presente en tejido sano, Flechas rojas TAD *de novo* en el tumor. Genes sobreexpresados en rojo, genes subexpresados en azul (Matrices a 50kb de resolución).

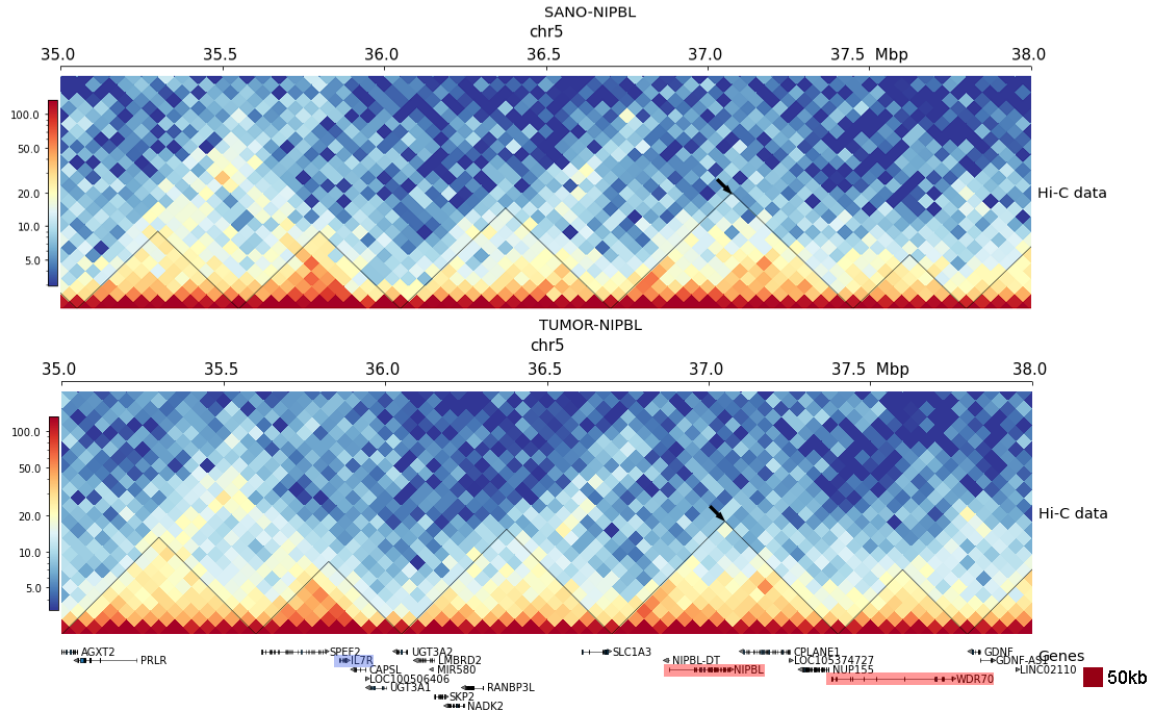


Figura 44. El gen de cargado de la cohesina a la cromatina *NIPBL* se localiza en un TAD equivalente entre el tejido adyacente sano y el tumor de CG difuso del paciente 8. Comparación de matrices de Hi-C entre tejido adyacente sano y tumor del locus que contiene al gen *NIPBL*. Flechas negras TAD presente en ambos tejidos. Genes sobreexpresados en rojo, genes subexpresados en azul (Matrices a 50kb de resolución).

Tomando estos datos en conjunto se muestra que los genes de regulación, cargado y descargado de la cohesina a la cromatina se encuentran en *locus* alterados topológicamente en el tumor del subtipo difuso del paciente 8 respecto al tejido adyacente sano, este hecho puede estar relacionado con la sobreexpresión de los mismos detectada por scRNA-seq y puede explicar el cambio topológico global en el tumor. Por lo tanto, se proponen estos *loci* para estudios más detallados en biopsias de CG difuso.

Discusión

El cáncer gástrico es un problema de salud relevante en México y a nivel mundial [1,2,11]. Existen pocas terapias eficientes para tratarlo y además las mismas no distinguen entre subtipos[3]. Aunque se han llevado a cabo esfuerzos para clasificarlo a nivel molecular, la heterogeneidad genética que presenta el CG hace difícil implementar las clasificaciones moleculares en la clínica[8]. En este contexto, conocer no solo a nivel codificante al CG sino también a nivel regulatorio podría mejorar nuestro entendimiento sobre el mismo y permitirnos tratarlo de manera dirigida y eficaz.

Como se mencionó a lo largo del texto, los cambios en la topología del genoma son un factor que puede impactar en la severidad de distintos tipos de cáncer, por ejemplo en el desarrollo de la metástasis[7,63]. En nuestro país; el presente proyecto es el primer acercamiento a este aspecto de la regulación genética en el cáncer gástrico. En particular, el cáncer gástrico del tipo difuso afecta a personas más jóvenes, es más agresivo y la sobrevida esperada es menor[3–5].

En un estudio epidemiológico del 2018 de la universidad de Texas se relaciona el CG del tipo difuso con la población hispana[67]. No obstante, en México no se tiene claro cual es la prevalencia de cada subtipo. El subtipo difuso en poblaciones mexicanas, se ha asociado a polimorfismos de nucleótido único (SNPs, por sus siglas en inglés) del gen CDH1[94]. Las mutaciones de dicho gen se han relacionado con alteraciones en la formación correcta de los epitelios gástricos y CG hereditario[5,8]. Sin embargo, no se ha indagado en la topología genómica del CG en pacientes mexicanos. Además, los trabajos de clasificación molecular realizados por el TCGA y el ACRG no consideran poblaciones hispanas[67]. Es por eso que es fundamental seguir investigando esta patología en nuestro país.

Los experimentos realizados en esta tesis en las PMBCs, tejido adyacente sano y tumor de cáncer gástrico difuso del paciente 8 muestran se estableció con éxito el protocolo de Hi-C en estas muestras y recuperamos información topológica relevante de biopsias que contienen menos de 5 millones de células.

Así mismo, se detectaron VE en datos de Hi-C con 88 millones de Unique Di-Tags, como se muestra en el análisis comparativo realizado con los datos generados por el grupo de Ooi *et al.* del tumor gástrico del subtipo intestinal T2000877. Dicho análisis confirma la sensibilidad que tiene la técnica y lo robustas que son las VE al ser detectadas en este tipo de datos genómicos. Por lo tanto, el Hi-C es óptimo para proyectos que busquen variaciones estructurales en muestras biológicas con muestras iguales o menores a 5 millones de células, en concreto biopsias de cáncer.

Las clasificaciones moleculares del CG relacionan al subtipo histológico intestinal con inestabilidad genómica y al subtipo difuso con estabilidad[5,8], cabe señalar que según los resultados del ACRG más del 80% de los tumores del subtipo difuso que analizaron son genómicamente estables[4]. Este dato muestra la importancia de buscar VE en ambos subtipos, ya que ambos pueden contener particularidades moleculares relevantes que permitan elegir un tratamiento por encima de otro, este punto se discutirá con más detalle adelante.

Los resultados encontrados en este trabajo son consecuentes con las clasificaciones del TCGA y el ACRG ya que el tumor gástrico del subtipo difuso del paciente 8 es genómicamente estable, Debido a que no se encontraron VE mayores a 1Mb, como se pueden observar en los datos de Hi-C del tumor de CG del subtipo intestinal T2000877, ni las amplificaciones características en las matrices de Hi-C que se presentan en la línea celular de CG SNU16. Ambos datos pertenecientes a CG genómicamente inestable, según las clasificaciones antes mencionadas[4,7,10,88] (Ver matrices de la figura 32).

Este trabajo es relevante en distintos niveles respecto a la regulación de la expresión genética en el CG del subtipo difuso, al integrar el componente topológico y transcripcional mediante Hi-C y scRNA-seq. Se encontró en primer lugar que cerca de la mitad de los TADs analizados (1667) entre el tejido adyacente sano y el tumor están presentes solo en este último. Mostrando un cambio global en las fronteras de los TADs si de considera que estas estructuras están conservadas entre tipos celulares[29]. Como resultado, existen interacciones distintas ente *loci* genómicos en el tumor respecto al tejido sano. Lo anterior genera un impacto

relevante en la expresión genética de estas células que presentan 1811 GDE en el tumor, de los cuales 717 GDE se encuentran dentro de 452 TADs exclusivos del tumor.

Al realizar la ontología de los GDE relacionándolos con procesos biológicos mediante la base de datos REACTOME[85], el de cargado de la cohesina a la cromatina es el proceso más relevante, tomando en cuenta el score combinado que arroja el análisis. Tal proceso es fundamental en la arquitectura del genoma por lo tanto se decidió profundizar en este resultado.

Las proteínas que forman parte de la cohesina y que regulan su asociación con la cromatina pueden estar relacionadas con el cambio topológico general que se presenta en el tumor al compararlo con el tejido adyacente sano, sobre todo teniendo en cuenta los reportes donde el silenciamiento de *NIPBL*, *WAPL* y *RAD21* tienen implicaciones globales en la topología y expresión genética [10,28,30]. De hecho, el único gen de este proceso que no se encuentra en un TAD diferencial del tumor es *NIPBL* el cual está en un TAD equivalente respecto al tejido sano. Sin embargo, *RAD21, SMC1A, SMC3,STAG2, WAPL* y *PDS5*A se encuentran en TADs exclusivos del tumor del paciente 8. Así mismo, estos 7 transcritos están sobreexpresados respecto al tejido adyacente sano y se observan cambios en la expresión de genes que se encuentran en el mismo vecindario genómico (Ver figuras 41 y 43), robusteciendo la hipótesis del impacto que tiene el cambio en la topología del genoma en la expresión genética del tumor.

De manera relevante, en 2015 un estudio de la universidad de Seúl reportó la sobre expresión en 16 tumores de CG de una muestra de 24 tumores, de las subunidades de la cohesina RAD21 y SMC1A cuya sobreexpresión se relacionó con peor pronóstico para los pacientes con esta malignidad[10]. Además, mediante ensayos de silenciamiento de ambos transcritos concluyeron que en la línea celular de CG genómicamente inestable SNU16 a lo largo de las divisiones celulares se disminuyó la amplificación en el genoma de oncogenes como CD44 que tiene un rol importante en la formación e identificación de células troncales de cáncer[14,95–97] y c-Myc el oncogén más amplificado en cáncer[25], aunado a una mayor sensibilidad de las

células silenciadas contra *RAD21* y *SMC1A* al tratamiento con cisplatino y el inhibidor de la poly ADP-ribosa (PARP por sus siglas en inglés). Es decir, fármacos que provocan daño al ADN promoviendo la apoptosis[10].

Esto es muestra del papel de *RAD21* y *SMC1A* en la inestabilidad genómica del cáncer gástrico. No obstante, en las líneas celulares de cáncer colorectal HCT116 y LoVo y la línea celular de cáncer hepático HepG2 las cuales son genómicamente estables, no presentaron detrimento en la amplificación de oncogenes ante el silenciamiento de estas proteínas. Además, se reportó una menor eficiencia de los efectos antiproliferativos del cisplatino y el inhibidor de PARP. Esto sugiere un papel dependiente de la estabilidad genómica en la actividad de estas subunidades de la cohesina en el cáncer y la resistencia a fármacos que provocan daño al ADN en tumores genómicamente estables.

Finalmente, el silenciamiento de *STAG2* en células HeLa resultó en inestabilidad genómica, esto se relaciona con la participación del complejo de la cohesina en la reparación homologa del rompimiento de la doble cadena de ADN reclutando maquinaria de reparación y manteniendo a la cromátida hermana rota cerca de la intacta que servirá de molde[28,29].

A pesar de la relación que se ha hecho entre la sobrexpresión de *RAD21* y *SMC1A* y la inestabilidad genómica en el CG es muy importante recalcar que siendo generalmente el subtipo difuso genómicamente estable la implicación de la sobreexpresión de todos estos transcritos no está clara aún y podría ser dependiente del subtipo tumoral. Es probable que en el cáncer gástrico del subtipo difuso la cohesina mantenga su papel de reparación de manera exacerbada considerando por ejemplo, la sobreexpresión de STAG2, haciendo resistente a este subtipo a tratamientos basados en agentes de daño al ADN como el cisplatino y los inhibidores de PARP.

Esta observación aporta nueva información para el uso de tratamientos adecuados en el CG difuso, puede que la regulación genética fina como la generación de nuevos contactos entre *loci* y compartimentos explique parte del proceso del cáncer en este subtipo tumoral y no de manera tan importante las grandes VE. No obstante,

cabe señalar que en tanto al conocimiento del cáncer gástrico en México y su regulación genética hacen falta estudios con más tumores para realizar generalizaciones.

Tomando toda esta evidencia en conjunto y contrastándola con los resultados obtenidos en este trabajo, se sugiere que la sobreexpresión de los factores involucrados en el cargado de la cohesina a la cromatina pueden tener un papel importante en la topología general del genoma del cáncer gástrico del tipo difuso impactando en su expresión genética de forma global. Por lo tanto los locus donde se mapean estos genes serían candidatos para más estudios en modelos de CG.

Conclusiones

1) Se estandarizó con éxito el protocolo Hi-C para detectar VE y TADs en PMBCs, tejido sano adyacente y cáncer gástrico difuso.

2) Los tejidos PMBCs, tejido sano adyacente y cáncer gástrico difuso analizados son genómicamente estables respecto a VE mayores o iguales a 1Mb.

3) Los datos de Hi-C son altamente sensibles para detectar VE mayores a 1MB.

4) Se encontró un cambio topológico a nivel de TADs y transcripcional general en el tumor de CG difuso en comparación con el tejido adyacente sano.

5) Los genes estructurales (*SMC1A*, *SMC3*, *RAD21* y *STAG2*) y relacionados con el cargado y mantenimiento de la cohesina en la cromatina (*WAPL* y *PDS5A*) están sobreexpresados en el tumor y se encuentran en TADs exclusivos del mismo, a excepción del gen de descargado de la cohesina a la cromatina *NIPBL*.

En conjunto, el cáncer gástrico del tipo difuso puede estar relacionado con alteraciones topológicas finas como la formación de nuevos contactos y no de manera tan importante con VE. Sin embargo, se requieren más datos para generalizar esta observación.

Perspectivas

Es necesario indagar más profundamente en otros aspectos de la topología en estos datos, por ejemplo en el componente de los compartimentos de cromatina ya que también puede haber alteraciones que expliquen las diferencias transcripcionales entre el tejido adyacente sano y el tumoral.

Además, la comparación y suma de otras matrices de Hi-C y datos de scRNA-seq de tumores de cáncer gástrico permitirán la generalización de las observaciones hechas en este trabajo y aumentar la resolución con la que se puede estudiar la topología del genoma y expresión genética del CG, así como las poblaciones celulares presentes en el mismo.

El análisis de identificación de potenciadores como la inmunoprecipitación de la cromatina contra las marcas de histonas H3K4me1 y H3K27ac permitirá conocer con mayor detalle la regulación genética de los *loci* propuestos.

Anexo de Tablas de Primers

Primers para adaptadores Tru-Seq Figuras 19,23 y 24	
Primer 1.0	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA
Primer 2.0	CAAGCAGAAGACGGCATACGAGAT

	Nombre	Secuencia5'-3'	Amplicon hasta DPNII o Mbo1	Tipo de control
Figura 18	C6-3	AGCAGCTGAAAACGGACTCGT	290	Externo
	Hist_f2	CATCCAGGGTATCACCAAGCC	120	
	TAD 5-6 H	GCCACTCGGCAGTCAGTATT	137	Interno
	TAD 5-6 M	GCTCGTCTGAGACCCTTCAC	49	
Figura 21	C6-2	CCCATTTTCGGCGTCGAGT	274	Externo
	C6-3	AGCAGCTGAAAACGGACTCGT	290	
	C6-3	AGCAGCTGAAAACGGACTCGT	290	Interno
	Hist_f3	GCTGCAGTAACAGTTCCGCCGT	229	

Anexo de Carta de consentimiento informado

INSTITUTO NACIONAL DE CIENCIAS MEDICAS Y NUTRICION SALVADOR ZUBIRAN

INST. NACIONAL DE CIENCIAS MÉDICAS Y NUTRICIÓN "SALVADOR ZUBIRÁN"
03 MAR 2022
COMITE DE ÉTICA EN INVESTIGACIÓN

CARTA DE CONSENTIMIENTO INFORMADO PARA PARTICIPAR EN EL PROYECTO: Identificación de subtipos celulares, vías moleculares y paisajes cromosómicos asociados con la progresión y malignidad tumoral.

Versión 3, 25 de febrero de 2022

Investigador principal: Dra. Yanin Chávarri Guerra
Dirección: Vasco de Quiroga 15, Sección XVI, Tlalpan, 14080, CDMX, México.
Teléfono de contacto de los investigadores: 54870900, Extensión 2254 y 5712.
Investigadores participantes: Dr. Enrique Soto Pérez de Celis, Dra Mayra Furlan Magaril, Dra. Paulina Licona-Limón, Ma. Aura Stephenson Gussinye, Dra. Blanca Ruiz Medina, Ma. Rosario Pérez Molina, Lic. Andrea Morales Alfaro
Nombre del patrocinador del estudio: Fondo Institucional de Fomento Regional para el Desarrollo Científico, Tecnológico y de Innovación
Dirección del patrocinador: Av. Insurgentes Sur 1582, Crédito Constructor, Benito Juárez, C.P 03940, CDMX.
Versión del consentimiento informado y fecha de su preparación: Versión 2, 17 de marzo de 2021

INTRODUCCIÓN:

Este documento es una invitación a participar en un estudio de investigación del Instituto. Por favor, tome todo el tiempo que sea necesario para leer este documento; pregunte al investigador sobre cualquier duda que tenga.
Este consentimiento informado cumple con los lineamientos establecidos en el Reglamento de la Ley General de Salud en Materia de Investigación para la salud, la Declaración de Helsinki y a las Buenas Prácticas Clínicas emitidas por la Comisión Nacional de Bioética.
Para decidir si participa o no en este estudio, usted debe tener el conocimiento suficiente acerca de los riesgos y beneficios con el fin tomar una decisión informada. Este formato de consentimiento informado le dará información detallada acerca del estudio de investigación que podrá comentar con su médico tratante o con algún miembro del equipo de investigadores. Al final se le pedirá que forme parte del proyecto y de ser así, bajo ninguna presión o intimidación, se le invitará a firmar este consentimiento informado.

Procedimiento para dar su consentimiento. Usted tiene el derecho a decidir si quiere participar o no como sujeto de investigación en este proyecto. El investigador le debe explicar ampliamente los beneficios y riesgos del proyecto sin ningún tipo de presión y **usted tendrá todo el tiempo que requiera para pensar, solo o con quien usted decida consultarlo, antes de decidir si acepta participar**. Cualquiera que sea su decisión no tendrá efecto alguno sobre su atención médica en el Instituto.
Con el fin de tomar una decisión verdaderamente informada sobre si acepta participar o

INSTITUTO NACIONAL DE CIENCIAS MEDICAS Y NUTRICION SALVADOR ZUBIRAN

no en este estudio, usted debe tener el conocimiento suficiente acerca de los posibles riesgos y beneficios a su salud al participar. Este documento le dará información detallada acerca del estudio de investigación, la cual podrá comentar quien usted quiera, por ejemplo, un familiar, su médico tratante, el investigador principal de este estudio o con algún miembro del equipo de investigadores. Al final, una vez leída y entendida esta información, se le invitará a que forme parte del proyecto y si usted acepta, sin ninguna presión o intimidación, se le invitará a firmar este consentimiento informado.
Este consentimiento informado cumple con los lineamientos establecidos en el Reglamento de la Ley General de Salud en Materia de Investigación para la Salud, la Declaración de Helsinki, y a las Buenas Prácticas Clínicas emitidas por la Comisión Nacional de Bioética.

Al final de la explicación, usted debe entender los puntos siguientes:

I. La justificación y los objetivos de la investigación.
II. Los procedimientos que se utilizarán y su propósito, incluyendo la identificación de qué son procedimientos experimentales.
III. Los riesgos o molestias previstos.
IV. Los beneficios que se pueden observar.
V. Los procedimientos alternativos que pudieran ser ventajosos para usted
VI. Garantía para recibir respuestas a las preguntas y aclarar cualquier duda sobre los procedimientos, riesgos, beneficios y otros asuntos relacionados con la investigación y el tratamiento de la materia.
VII. La libertad que tiene de retirar su consentimiento en cualquier momento y dejar de participar en el estudio, sin que por ello se afecte su atención y el tratamiento en el Instituto.
VIII. La seguridad de que no se le va a identificar de forma particular y que se mantendrá la confidencialidad de la información relativa a su privacidad.
IX. El compromiso del investigador de proporcionarle la información actualizada que pueda ser obtenida durante el estudio, aunque esto pudiera afectar a su disposición para continuar con su participación.
X. La disponibilidad del tratamiento médico y compensación a que legalmente tiene derecho, en el caso de que ocurran daños causados directamente por la investigación.

Puede solicitar más tiempo o llevar a casa este formulario antes de tomar una decisión final en los días futuros.

INVITACION A PARTICIPAR COMO SUJETO DE INVESTIGACION Y DESCRIPCIÓN DEL PROYECTO

Estimado(a) Sr(a).

El Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ) y el Instituto de Fisiología Molecular de la Universidad Nacional Autónoma de México, a través

INSTITUTO NACIONAL DE CIENCIAS MEDICAS Y NUTRICION SALVADOR ZUBIRAN

de un grupo de investigación, le invitan a participar como sujeto de investigación en este estudio que tiene como **objetivo:** detectar y analizar anormalidades en el material genético (es la información que contienen las células del cuerpo y que le indican como funcionar, crecer y desarrollarse) en muestras de pacientes mexicanos con diagnóstico de cáncer de mama y cáncer gástrico, a través del uso de técnicas de laboratorio.
La duración total del estudio es de tres años. El número aproximado de participantes que se incluirán en este estudio serán de 40 pacientes (20 pacientes con diagnóstico de cáncer de seno y 20 pacientes con sospecha o diagnóstico de cáncer de estómago)
Usted fue invitado al estudio debido a que tiene las siguientes características: ser mayor de edad (18 años), tiene diagnóstico de cáncer de seno o sospecha/diagnóstico de cáncer de estómago, será sometido a algún procedimiento programado de rutina para toma de muestra (biopsia, cirugía o endoscopia), no ha recibido tratamiento con quimioterapia y no tiene ninguna enfermedad o tratamiento que cause inmunosupresión (baja de defensas) o alguna infección activa.

PROCEDIMIENTOS DEL ESTUDIO

Su participación en el estudio consiste en: a) obtención de una muestra de sangre y b) la toma de muestras de tejido.

INSTITUTO NACIONAL DE CIENCIAS MEDICAS Y NUTRICIÓN "SALVADOR ZUBIRÁN" 03 MAR 2020 COMITÉ DE ÉTICA EN INVESTIGACIÓN

a) Toma de muestra de sangre. Una vez que usted haya aceptado participar en el protocolo y firmado el consentimiento informado se le tomará una muestra de sangre de 3 ml aprox. (una cucharadita) en un tubo con tapa morada y se mantendrá en refrigeración.
b) Las muestras de tejido. Se obtendrán cuando usted tenga algún procedimiento programado de rutina para toma de muestra, como cirugía, endoscopía o biopsias de rutina.
Posterior a esto su participación habrá concluido.
Sus muestras se enviarán al laboratorio de Fisiología Celular de la UNAM para su análisis y al laboratorio de patología del INCMNSZ.

Las intervenciones propuestas que son experimentales: hasta el momento nuestro estudio es considerado como experimental, por lo que usted no recibirá resultados de dicho estudio.
Las intervenciones incluidas en el estudio que son parte de su tratamiento estándar son: La valoración por parte de su oncólogo y cualquier otro médico del instituto no tendrá cambios.
Las responsabilidades de los participantes incluyen: Proporcionar una muestra de sangre y tumor.

RIESGOS E INCONVENIENTES

Los datos acerca de su identidad, su información médica, y la identidad de sus familiares no serán revelados en ningún momento como lo estipula la ley, por tanto, en la recolección de datos clínicos usted no enfrenta riesgos mayores a los relativos a la protección de la confidencialidad, la cual será protegida mediante la eliminación de cualquier dato personal que lo relacione con las muestras.
Los riesgos relacionados con la extracción de sangre son mínimos, entre ellos, moretones en el lugar de la extracción, necesidad de repetir la punción, y dolor temporal.
Los riesgos relacionados con la obtención de una muestra de tumor son moretones en el lugar de la biopsia, infecciones y dolor temporal para el caso del cáncer de mama y la

INSTITUTO NACIONAL DE CIENCIAS MEDICAS Y NUTRICION SALVADOR ZUBIRAN

obtención de muestra por medio de endoscopia puede presentarse sangrado de tubo digestivo, desgarro del tracto gastrointestinal y reacción a la sedación. Estos riesgos no aumentarán con la obtención de la muestra que nosotros tomaremos para la investigación y son propios del procedimiento que de cualquier forma le realizarán al paciente para diagnóstico del tumor.

BENEFICIOS POTENCIALES

Este estudio no está diseñado para beneficiarle directamente. Sin embargo, el conocer mejor los mecanismos del cáncer en la población mexicana permitirá en un futuro diseñar mejores herramientas de diagnóstico y tratamiento. Por lo tanto, su participación altruista podría beneficiar a otros pacientes con cáncer de mama triple negativo y gástrico.

CONSIDERACIONES ECONÓMICAS

No se cobrará ninguna tarifa por participar en el estudio ni se le hará pago alguno. Coordinaremos la visita de este proyecto con sus visitas estándar al hospital para no incurrir en ningún gasto extra.

COMPENSACION

Este estudio no acarrea riesgos adicionales para su salud más que los considerados por el procedimiento de diagnóstico o tratamiento de cáncer de mama y cáncer gástrico que los médicos encargados de su cuidado han considerado para usted. Los costos que se generen por la atención de complicaciones generadas durante el procedimiento de toma de muestra no serán cubiertos por nosotros. El Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán no brinda ningún tipo adicional de compensación para cubrir el daño.

ALTERNATIVAS A SU PARTICIPACIÓN:

Su participación es voluntaria. Por lo que usted puede elegir no participar en el estudio. En caso de decidir no participar, usted seguirá recibiendo el tratamiento o manejo habitual (estándar) para su enfermedad.

POSIBLES PRODUCTOS COMERCIALES DERIVABLES DEL ESTUDIO:

Si un producto comercial es desarrollado como resultado del estudio, tal insumo será propiedad del Instituto Nacional de Ciencias Médicas y Nutrición, Salvador Zubirán (INCMNSZ), del Instituto de Fisiología Celular de la UNAM o quienes ellos designen. En tal caso, usted no recibirá un beneficio financiero por el mismo.

PARTICIPACIÓN Y RETIRO DEL ESTUDIO:

Recuerde que su participación es VOLUNTARIA. Si usted decide no participar, tanto su relación habitual con el INCMNSZ como su derecho para recibir atención médica o cualquier servicio al que tenga derecho no se verán afectados. Si decide participar, tiene la libertad para retirar su consentimiento e interrumpir su participación en cualquier momento sin perjudicar su atención en el INCMNSZ. En ese caso le pediremos que se lo comunique a la Dra. Yanin Chávarri Guerra o al Dr. Enrique Soto Pérez de Celis

El investigador o el patrocinador del estudio pueden **excluirlo del estudio si:** no obtenemos la muestra requerida para este estudio.

INSTITUTO NACIONAL DE CIENCIAS MÉDICAS Y NUTRICIÓN "SALVADOR ZUBIRÁN"
03 MAR 2022
COMITÉ DE ÉTICA EN INVESTIGACIÓN

INSTITUTO NACIONAL DE CIENCIAS MEDICAS Y NUTRICION SALVADOR ZUBIRAN

CONFIDENCIALIDAD Y MANEJO DE SU INFORMACIÓN

Su nombre no será usado en ningún momento. Su confidencialidad será protegida como lo marca la ley, asignando códigos a su información. El código es un número de identificación que no incluye datos personales por ello su información será mantenida a través de la asignación de códigos. Ninguna información sobre su persona será compartida con otros sin su autorización, excepto:

- Si es necesario para proteger sus derechos y bienestar (por ejemplo, si ha sufrido una lesión y requiere tratamiento de emergencia); o
- Si es solicitado por la ley.

Monitores o auditores del estudio podrán tener acceso a la información de los participantes.

Si usted decide retirarse del estudio, podrá solicitar el retiro y destrucción de su información y muestra almacenada.

Los Comités de investigación y de Ética en Investigación del Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán aprobaron la realización de este estudio. Dichos comités revisan, aprueban y supervisan los estudios de investigación en humanos en el Instituto.

Los datos científicos obtenidos como parte de este estudio podrían ser utilizados en publicaciones o presentaciones médicas. Su nombre y otra información personal serán eliminados antes de usar los datos.

Si usted lo solicita su médico de cabecera será informado sobre su participación en el estudio.

Su nombre no será usado en ninguno de los reportes públicos del estudio. Sus datos no podrán ser usados para estudios de investigación que estén relacionados con condiciones distintas a las estudiadas en este proyecto, y estos estudios deberán ser sometidos a aprobación por un Comité de Ética.

Si bien existe la posibilidad de que su privacidad sea afectada como resultado de su participación en el estudio, su confidencialidad será protegida como lo marca la ley, asignando códigos a su información. El código es un número de identificación que no incluye datos personales. Ninguna información sobre su persona será compartida con otros sin su autorización, excepto:

- Si es necesario para proteger sus derechos y bienestar (por ejemplo, si ha sufrido una lesión y requiere tratamiento de emergencia); o
- Es solicitado por la ley.

Si usted decide retirarse del estudio, podrá solicitar el retiro y destrucción de su información.

Riesgos asociados con el incumplimiento del requisito de confidencialidad: Existe un leve riesgo de que personas no vinculadas a este estudio accidentalmente descubran su identidad u obtengan su información personal.

IDENTIFICACIÓN DE LOS INVESTIGADORES:

Si usted tiene preguntas sobre el estudio, puede ponerse en contacto con la Dra. Yanin Chávarri Guerra en el INCMNSZ al teléfono 54870900 Ext 2254 o con el Dr. Enrique Soto Pérez de Celis en el INCMNSZ al teléfono 54870900 Ext 5712.

En caso de que usted sufra un daño relacionado al estudio, por favor póngase en contacto con la Dra. Yanin Chávarri Guerra en el INCMNSZ al teléfono 54870900 Ext 2254 o 044554128608 y con el Dr. Enrique Soto Pérez de Celis en el INCMNSZ al teléfono 54870900.

INSTITUTO NACIONAL DE CIENCIAS MÉDICAS Y NUTRICIÓN "SALVADOR ZUBIRÁN"
03 MAR 2022
COMITÉ DE ÉTICA EN INVESTIGACIÓN

INSTITUTO NACIONAL DE CIENCIAS MEDICAS Y NUTRICION SALVADOR ZUBIRAN

Si usted tiene preguntas acerca de sus derechos como participante en el estudio, puede hablar con el Presidente del Comité de Ética en Investigación del INCMNSZ (Arturo Galindo Fraga, tel: 54870900. ext. 6101).

DECLARACIÓN DEL CONSENTIMIENTO INFORMADO

He leído con cuidado este consentimiento informado, he hecho todas las preguntas que he tenido y todas han sido respondidas satisfactoriamente. Para poder participar en el estudio, estoy de acuerdo con todos los siguientes puntos:
Estoy de acuerdo en participar en el estudio descrito anteriormente. Los objetivos generales, particulares del reclutamiento y los posibles daños e inconvenientes me han sido explicados a mi entera satisfacción.
Estoy de acuerdo, en caso de ser necesario, que se me contacte en el futuro si el proyecto requiere colectar información adicional o si encuentran información relevante para mi salud.
Mi firma también indica que he recibido un duplicado de este consentimiento informado.

Por favor responda las siguientes preguntas:

		SÍ (marque por favor)	NO (marque por favor)
a.	¿Ha leído y entendido el formato de consentimiento informado, en su lengua materna?	☐	☐
b.	¿Ha tenido la oportunidad de hacer preguntas y de discutir este estudio?	☐	☐
c.	¿Ha recibido usted respuestas satisfactorias a todas sus preguntas?	☐	☐
d.	¿Ha recibido suficiente información acerca del estudio y ha tenido el tiempo suficiente para tomar la decisión?	☐	☐
e.	¿Entiende usted que su participación es voluntaria y que es libre de suspender su participación en este estudio en cualquier momento sin tener que justificar su decisión y sin que esto afecte su atención médica o sin la pérdida de los beneficios a los que de otra forma tenga derecho?	☐	☐
f.	¿Entiende los posibles riesgos, algunos de los cuales son aún desconocidos, de participar en este estudio?	☐	☐
g.	¿Entiende que puede no recibir algún beneficio directo de participar en este estudio?	☐	☐
h.	¿Entiende que no está renunciando a ninguno de sus derechos legales a los que es acreedor de otra forma como sujeto en un estudio de investigación?	☐	☐
i.	¿Entiende que el médico participante en el estudio puede retirarlo del mismo sin su consentimiento, ya sea debido a que	☐	☐

INSTITUTO NACIONAL DE CIENCIAS MEDICAS Y NUTRICION SALVADOR ZUBIRAN

		SÍ (marque por favor)	NO (marque por favor)
	usted no siguió los requerimientos del estudio o si el médico participante en el estudio considera que médicamente su retiro es en su mejor interés?		
j.	¿Entiende que usted recibirá un original firmado y fechado de esta Forma de Consentimiento para sus registros personales?	☐	☐

Declaración del paciente:

Yo,_____ _____

declaro que es mi decisión participar como sujeto de investigación en el estudio. Mi participación es voluntaria. He sido informado que puedo negarme a participar o terminar mi participación en cualquier momento del estudio sin que sufra penalidad alguna o pérdida de beneficios. Si suspendo mi participación, recibiré el tratamiento médico habitual al que tengo derecho en el Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ) y no sufriré perjuicio en mi atención médica ni en futuros estudios de investigación. Yo puedo solicitar información adicional acerca de los riesgos o beneficios potenciales derivados de mi participación en este estudio.

Puedo obtener los resultados de mis exámenes clínicos si los solicito. Si tengo preguntas sobre el estudio, puedo ponerme en contacto con la Dra. Yanin Chávarri Guerra al teléfono 54870900 EXT 2254. Si tengo preguntas sobre mis derechos como participante en el estudio, problemas, preocupaciones o preguntas, si quiero obtener información, u ofrecer información que puede hablar con el presidente del Comité de Ética de Investigación de INCMNSZ (Dr. Arturo Galindo Fraga Tel: 54870900. ext 6101). Debo informar a los investigadores de cualquier cambio en mi estado de salud (por ejemplo, uso de nuevos medicamentos, cambios en el consumo de tabaco) o en la ciudad donde resido, tan pronto como sea posible. He leído y entendido toda la información que me han dado sobre mi participación en el estudio. He tenido la oportunidad para discutirlo y hacer preguntas. Todas las preguntas han sido respondidas a mi satisfacción. He entendido que recibiré una copia firmada de este consentimiento informado.

INSTITUTO NACIONAL DE CIENCIAS MÉDICAS Y NUTRICIÓN "SALVADOR ZUBIRÁN" 03 MAR 2015 COMITÉ DE ÉTICA EN INVESTIGACIÓN

Nombre del Participante	Firma del Participante	Fecha

Coloque su huella digital si no sabe escribir

Nombre del representante legal (si aplica)	Firma del representante legal	Fecha

INSTITUTO NACIONAL DE CIENCIAS MEDICAS Y NUTRICION SALVADOR ZUBIRAN

Nombre del Investigador que explicó el documento	Firma del Investigador	Fecha
Nombre del Testigo 1	Firma del Testigo 1	Fecha

Relación con el participante:

Dirección:

Nombre del Testigo 2	Firma del Testigo 2	Fecha

Dirección:

Relación que guarda con el participante:

Lugar y Fecha:

(El presente documento es original y consta de 8 páginas)

INST. NACIONAL DE CIENCIAS MEDICAS Y NUTRICION "SALVADOR ZUBIRÁN"
03 MAR 2020
COMITE DE ÉTICA EN INVESTIGACIÓN

Bibliografía

1. Sampieri, C. L. & Mora, M. Gastric cancer research in Mexico: A public health priority. *World J Gastroenterol* (2014) doi:10.3748/wjg.v20.i16.4491.

2. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* (2018) doi:10.3322/caac.21492.

3. Orditura, M. *et al.* Treatment of gastric cancer. *World J Gastroenterol* **20**, (2014).

4. Cristescu, R. *et al.* Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat Med* **21**, (2015).

5. Sitarz, R. *et al.* Gastric cancer: Epidemiology, prevention, classification, and treatment. *Cancer Management and Research* Preprint at https://doi.org/10.2147/CMAR.S149619 (2018).

6. Rawla, P. & Barsouk, A. Epidemiology of gastric cancer: Global trends, risk factors and prevention. *Przeglad Gastroenterologiczny* Preprint at https://doi.org/10.5114/pg.2018.80001 (2019).

7. Ooi, W. F. *et al.* Integrated paired-end enhancer profiling and whole-genome sequencing reveals recurrent CCNE1 and IGF2 enhancer hijacking in primary gastric adenocarcinoma. *Gut* **69**, 1039–1052 (2020).

8. Chivu-Economescu, M. *et al.* New therapeutic options opened by the molecular classification of Gastric cancer. *World Journal of Gastroenterology* https://doi.org/10.3748/wjg.v24.i18.1942 (2018).

9. Bass, A. J. *et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* (2014) doi:10.1038/nature13480.

10. Yun, J. *et al.* Reduced cohesin destabilizes high-level gene amplification by disrupting pre-replication complex bindings in human cancers with chromosomal instability. *Nucleic Acids Res* **44**, 558–572 (2016).

11. CansecoEpidemiología de cáncer gástrico en el tercer nivel de atención en salud en Chiapas-Ávila, L. M. *et al.* Epidemiología de cáncer gástrico en el tercer nivel de atención en salud en Chiapas. *Rev Gastroenterol Mex* (2019) doi:10.1016/j.rgmx.2018.06.006.

12. Martínez-Galindo, M. G. *et al.* Histopathologic characteristics of gastric adenocarcinoma in Mexican patients: A 10-year experience at the Hospital Juárez de México. *Revista de Gastroenterología de México (English Edition)* **80**, (2015).

13. Bandhavkar, S. Cancer stem cells: A metastasizing menace! *Cancer Medicine* https://doi.org/10.1002/cam4.629 (2016).

14. Medrano-Gonzálezl, P. A., Cruz-Villegas, F., Alarcón del Carmen, A., Montaño, L. F. & Rendón-Huerta, E. P. Claudin-6 increases SNAI1, NANOG and SOX2 gene expression in human gastric adenocarcinoma AGS cells. *Mol Biol Rep* **49**, (2022).

15. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nature Reviews Genetics* Preprint at https://doi.org/10.1038/nrg.2016.112 (2016).

16. Harewood, L. *et al.* Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol* (2017) doi:10.1186/s13059-017-1253-8.

17. Zhang, Y. *et al.* Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* (2012) doi:10.1016/j.cell.2012.02.002.

18. Kim, K., Eom, J. & Jung, I. Characterization of Structural Variations in the Context of 3D Chromatin Structure. *Mol Cells* (2019) doi:10.14348/molcells.2019.0137.

19. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (1979)* **326**, (2009).

20. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, (2014).

21. Pierro, M. di, Potoyan, D. A., Wolynes, P. G. & Onuchic, J. N. Anomalous diffusion, spatial coherence, and viscoelasticity from the energy landscape of human chromosomes. *Proc Natl Acad Sci U S A* **115**, (2018).

22. Ghosh, R. P. *et al.* A fluorogenic array for temporally unlimited single-molecule tracking. *Nat Chem Biol* **15**, (2019).

23. Ji, X. *et al.* 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell* **18**, (2016).

24. Ghosh, R. P. & Meyer, B. J. Spatial Organization of Chromatin: Emergence of Chromatin Structure during Development. *Annual Review of Cell and Developmental Biology* vol. 37 https://doi.org/10.1146/annurev-cellbio-032321-035734 (2021).

25. Hnisz, D., Schuijers, J., Li, C. H. & Young, R. A. Regulation and Dysregulation of Chromosome Structure in Cancer. *Annual Review of Cancer Biology* vol. 2 Preprint at https://doi.org/10.1146/annurev-cancerbio-030617-050134 (2018).

26. Pękowska, A. *et al.* Gain of CTCF-Anchored Chromatin Loops Marks the Exit from Naive Pluripotency. *Cell Syst* **7**, (2018).

27. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, (2015).

28. Di Nardo, M., Pallotta, M. M. & Musio, A. The multifaceted roles of cohesin in cancer. *Journal of Experimental and Clinical Cancer Research* vol. 41 Preprint at https://doi.org/10.1186/s13046-022-02321-5 (2022).

29. Losada, A. Cohesin in cancer: Chromosome segregation and beyond. *Nat Rev Cancer* **14**, (2014).

30. Davidson, I. F. & Peters, J. M. Genome folding through loop extrusion by SMC complexes. *Nature Reviews Molecular Cell Biology* vol. 22 Preprint at https://doi.org/10.1038/s41580-021-00349-7 (2021).

31. Luppino, J. M. *et al.* NIPBL and WAPL balance cohesin activity to regulate chromatin folding and gene expression. *bioRxiv* (2022).

32. Panigrahi, A. & O'Malley, B. W. Mechanisms of enhancer action: the known and the unknown. *Genome Biology* vol. 22 Preprint at https://doi.org/10.1186/s13059-021-02322-1 (2021).

33. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**, (2003).

34. Chepelev, I., Wei, G., Wangsa, D., Tang, Q. & Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res* **22**, (2012).

35. Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* vol. 461 Preprint at https://doi.org/10.1038/nature08451 (2009).

36. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, (2013).

37. Pott, S. & Lieb, J. D. What are super-enhancers? *Nature Genetics* vol. 47 Preprint at https://doi.org/10.1038/ng.3167 (2015).

38. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, (2012).

39. Erokhin, M., Vassetzky, Y., Georgiev, P. & Chetverina, D. Eukaryotic enhancers: Common features, regulation, and participation in diseases. *Cellular and Molecular Life Sciences* vol. 72 Preprint at https://doi.org/10.1007/s00018-015-1871-9 (2015).

40. Palstra, R. J. *et al.* The β-globin nuclear compartment in development and erythroid differentiation. *Nature Genetics* vol. 35 Preprint at https://doi.org/10.1038/ng1244 (2003).

41. Schoenfelder, S. *et al.* Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat Genet* **47**, (2015).

42. Denholtz, M. *et al.* Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and polycomb proteins in genome organization. *Cell Stem Cell* **13**, (2013).

43. Bantignies, F. *et al.* Polycomb-dependent regulatory contacts between distant hox loci in drosophila. *Cell* **144**, (2011).

44. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* (2012) doi:10.1038/nature11082.

45. Gong, Y. *et al.* Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nat Commun* **9**, (2018).

46. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, (2016).

47. Ghavi-Helm, Y. *et al.* Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat Genet* **51**, (2019).

48. Mirny, L. A., Imakaev, M. & Abdennur, N. Two major mechanisms of chromosome organization. *Current Opinion in Cell Biology* vol. 58 Preprint at https://doi.org/10.1016/j.ceb.2019.05.001 (2019).

49. van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* vol. 169 Preprint at https://doi.org/10.1016/j.cell.2017.04.022 (2017).

50. Pombo, A. & Dillon, N. Three-dimensional genome architecture: Players and mechanisms. *Nature Reviews Molecular Cell Biology* vol. 16 Preprint at https://doi.org/10.1038/nrm3965 (2015).

51. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harbor perspectives in biology* vol. 2 Preprint at https://doi.org/10.1101/cshperspect.a003889 (2010).

52. Rowley, J. D., le Beau, M. M. & Rabbitts, T. H. *Chromosomal translocations and genome rearrangements in cancer*. *Chromosomal Translocations and Genome Rearrangements in Cancer* (2015). doi:10.1007/978-3-319-19983-2.

53. *Holland-Frei Cancer Medicine*. *Holland-Frei Cancer Medicine* (2016). doi:10.1002/9781119000822.

54. Azad, N., Zahnow, C. A., Rudin, C. M. & Baylin, S. B. The future of epigenetic therapy in solid tumours - Lessons from the past. *Nature Reviews Clinical Oncology* vol. 10 Preprint at https://doi.org/10.1038/nrclinonc.2013.42 (2013).

55. Easwaran, H., Tsai, H. C. & Baylin, S. B. Cancer Epigenetics: Tumor Heterogeneity, Plasticity of Stem-like States, and Drug Resistance. *Molecular Cell* vol. 54 Preprint at https://doi.org/10.1016/j.molcel.2014.05.015 (2014).

56. You, J. S. & Jones, P. A. Cancer Genetics and Epigenetics: Two Sides of the Same Coin? *Cancer Cell* vol. 22 Preprint at https://doi.org/10.1016/j.ccr.2012.06.008 (2012).

57. Harewood, L. & Fraser, P. The impact of chromosomal rearrangements on regulation of gene expression. *Hum Mol Genet* (2014) doi:10.1093/hmg/ddu278.

58. Nowell, P. & Hungerford, D. A minute chromosome in human chronic 9 granulocytic leukemia. *Science (1979)* **132**, (1960).

59. Yao, F. *et al.* Recurrent Fusion Genes in Gastric Cancer: CLDN18-ARHGAP26 Induces Loss of Epithelial Integrity. *Cell Rep* **12**, (2015).

60. Tanaka, A. *et al.* Frequent CLDN18-ARHGAP fusion in highly metastatic diffusetype gastric cancer with relatively early onset. *Oncotarget* **9**, (2018).

61. Fournier, A. *et al.* 1q12 chromosome translocations form aberrant heterochromatic foci associated with changes in nuclear architecture and gene expression in B cell lymphoma. *EMBO Mol Med* **2**, (2010).

62. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, (2016).

63. Ren, B. *et al.* High-resolution Hi-C maps highlight multiscale 3D epigenome reprogramming during pancreatic cancer metastasis. *J Hematol Oncol* **14**, (2021).

64. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, (2017).

65. Sun, K. *et al.* scRNA-seq of gastric tumor shows complex intercellular interaction with an alternative T cell exhaustion trajectory. *Nat Commun* **13**, (2022).

66. Chen, Y. *et al.* Reconstruction of the gastric cancer microenvironment after neoadjuvant chemotherapy by longitudinal single-cell sequencing. *J Transl Med* **20**, (2022).

67. Sanjeevaiah, A., Cheedella, N., Hester, C. & Porembka, M. R. Gastric cancer: Recent molecular classification advances, racial disparity, and management implications. *Journal of Oncology Practice* vol. 14 Preprint at https://doi.org/10.1200/JOP.17.00025 (2018).

68. Wingett, S. *et al.* HiCUP: Pipeline for mapping and processing Hi-C data. *F1000Res* **4**, (2015).

69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, (2012).

70. Wang, S. *et al.* HiNT: A computational method for detecting copy number variations and translocations from Hi-C data. *Genome Biol* (2020) doi:10.1186/s13059-020-01986-5.

71. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, (2009).

72. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, (2016).

73. Xi, R., Lee, S., Xia, Y., Kim, T. M. & Park, P. J. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res* **44**, (2016).

74. Wolff, J. *et al.* Galaxy HiCExplorer 3: A web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res* **48**, (2020).

75. Ramírez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* **9**, (2018).

76. van der Maaten, L. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research* **15**, (2015).

77. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, (2021).

78. Bischoff, P. *et al.* Single-cell RNA sequencing reveals distinct tumor microenvironmental patterns in lung adenocarcinoma. *Oncogene* **40**, (2021).

79. Wu, F. *et al.* Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nat Commun* **12**, (2021).

80. He, S. *et al.* Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol* **21**, (2020).

81. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**, (2015).

82. Chen, E. Y. *et al.* Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, (2013).

83. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, (2016).

84. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr Protoc* **1**, (2021).

85. Gillespie, M. *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res* **50**, (2022).

86. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, (2020).

87. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res* **12**, (2002).

88. Ku, J.-L. & Park, J.-G. Biology of SNU Cell Lines. *Cancer Res Treat* **37**, (2005).

89. Ramírez, J. D. *et al.* The GALNTL6 Gene rs558129 Polymorphism Is Associated With Power Performance. *J Strength Cond Res* **34**, (2020).

90. Zhang, X. H. *et al.* A Review of Progress in Histone Deacetylase 6 Inhibitors Research: Structural Specificity and Functional Diversity. *J Med Chem* **64**, (2021).

91. Ooi, W. F. *et al.* Integrated paired-end enhancer profiling and whole-genome sequencing reveals recurrent CCNE1 and IGF2 enhancer hijacking in primary gastric adenocarcinoma. *Gut* (2020) doi:10.1136/gutjnl-2018-317612.

92. Kueng, S. *et al.* Wapl Controls the Dynamic Association of Cohesin with Chromatin. *Cell* **127**, (2006).

93. Pié, J. *et al.* Mutations and variants in the cohesion factor genes NIPBL, SMC1A, and SMC3 in a cohort of 30 unrelated patients with Cornelia de Lange syndrome. *Am J Med Genet A* **152**, (2010).

94. Bustos-Carpinteyro, A. R. *et al.* CDH1 somatic alterations in Mexican patients with diffuse and mixed sporadic gastric cancer. *BMC Cancer* **19**, (2019).

95. Huang, C. *et al.* ERK1/2-Nanog signaling pathway enhances CD44(+) cancer stem-like cell phenotypes and epithelial-to-mesenchymal transition in head and neck squamous cell carcinomas. *Cell Death Dis* **11**, (2020).

96. Senbanjo, L. T. & Chellaiah, M. A. CD44: A multifunctional cell surface adhesion receptor is a regulator of progression and metastasis of cancer cells. *Frontiers in Cell and Developmental Biology* vol. 5 Preprint at https://doi.org/10.3389/fcell.2017.00018 (2017).

97. Wang, L., Zuo, X., Xie, K. & Wei, D. The role of CD44 and cancer stem cells. in *Methods in Molecular Biology* (2018). doi:10.1007/978-1-4939-7401-6_3.